Humboldt-Universität zu Berlin – Geographisches Institut

# Visual Analytics for Detection and Assessment of Process-Related Patterns in Geoscientific Spatiotemporal Data

## Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)

im Fach Geographie

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
Dipl.-Geogr. Patrick Köthur

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät
Prof. Dr. Elmar Kulke

Gutachter/Gutachterinnen:
Prof. Dr. Doris Dransch
Prof. Dr. Liqiu Meng
Univ.-Prof. Torsten Möller, PhD

Eingereicht: 13. Juli 2015
Tag der Verteidigung: 14. Dezember 2015

# Acknowledgements

This thesis would not have been possible without the many people who supported me in the process. First and foremost, I would like to express my gratitude to my supervisor Doris Dransch. She gave me the freedom to follow my scientific interests and, most importantly, always provided guidance, support, and encouragement. Many thanks to Liqiu Meng and Torsten Möller for not hesitating to act as referees for this thesis.

This thesis is the result of highly interdisciplinary research and several collaborations with geoscientists. I would like to thank all my coauthors from the domain-expert side, Henryk Dobslaw, Julian Kuhlmann, Norbert Marwan, and Stefan Schinkel, for their open-mindedness, their enthusiasm, and, of course, for their patience in countless meetings and discussions.

I cannot thank my current and former colleagues (and sometimes coauthors) from the Geoinformatics Section of the German Research Centre for Geosciences GFZ enough for their unwavering support. I thank Mike Sips and Andrea Unger for providing so many inspiring ideas, feedback, and much-needed guidance, as well as for sharing so much of their time and invaluable experience with me. I also thank Carl Witt, Joachim Fohringer, Ralf Friedeman, Alexander Bobach, Tobias Rawald, and Janis Jatnieks for their support in implementing prototypes, for their important contributions to often spontaneous brainstorming sessions, for always listening, and for generally being great colleagues.

I would also like to thank Alan and Ajna, Heiko, Marek, and Anja for incredible friendships that have already endured decades and will hopefully last a lifetime. Furthermore, I would like to express my sincere gratitude to Silvia for her support. Lastly, a very special thanks to my grandparents, Rita and Lothar, who took me in at an early age and since then gave me nothing but love, security, and confidence.

# Abstract

The geosciences study the manifold processes within the complex Earth system to gain a better understanding of our planet. In order to obtain information about processes, spatiotemporal data is collected via observations or simulations. In these data, the spatial and temporal behavior of processes of the Earth system manifests itself as specific patterns. To gain insight into the various processes, scientists must detect and interpret these patterns by considering the geospatial and temporal variability in the data. This involves several challenges: (a) large spatiotemporal data spaces have to be taken into account, (b) only little aggregation and dimensionality reduction techniques should be applied to reduce loss of information, (c) prominent spatiotemporal patterns must be detected, and (d) among these patterns the ones that are related to processes within the Earth system must be identified. The overall aim of this thesis was to study how visual analytics, which combines automated analysis and interactive visual exploration, can address these challenges and facilitate the analysis of processes in geoscientific spatiotemporal data. To this end, this thesis focused on three important analysis perspectives and addressed each one in a separate research question: (1) How can visual analytics support the detection and assessment of prominent types of spatial situations? (2) How can visual analytics support the detection and assessment of prominent types of temporal behavior? (3) How can visual analytics improve the analysis of interrelations of temporal behavior? Three novel visual analytics solutions were developed, one to investigate each research question. The first solution helps geoscientists to analyze prominent spatial situations in the data and their occurrence over time. Interactive visual summaries were introduced as a means to facilitate this task. Hierarchical clustering is used to arrange all spatial situations in the data in a hierarchy of clusters. The combination with interactive visual analysis enables users to explore and alter the resulting hierarchy, to extract different sets of representative spatial situations, and to interpret and assess the corresponding spatiotemporal patterns. The second visual analytics solution supports geoscientists in the analysis of prominent types of temporal behavior and their location in geographic space. Cluster ensembles are integrated with interactive visual exploration to enable users to systematically detect and interpret various types of temporal behavior in different data sets and to use this information for assessment of simulation model output. The third solution enables geoscientists to detect and analyze interrelations of temporal behavior in the data. An established correlation-based technique for detection of interrelations between two individual time series (windowed cross-correlation) was extended to the comparison of entire ensembles of time series through visual analytics. It not only allows scientists to study interrelations, but also to assess how much these interrelations vary between two ensembles. The visual analytics solutions were developed following a rigorous user- and task-centered methodology and successfully applied to use cases in Earth system modeling, ocean modeling, paleoclimatology, and even cognitive science. The results of this thesis demonstrate that visual analytics successfully addresses the challenges outlined above. This thesis also demonstrates that visual analytics is a valuable approach to the analysis of process-related patterns in spatiotemporal data for at least three reasons: (1) geoscientific applications greatly benefit from the interactivity of visual analytics solutions because it supports a free-flowing analytical discourse with the data, (2) visual analytics is able to extend the boundaries of existing analysis methods, and (3) it introduces geoscientists to new, insightful perspectives on the data and the processes they describe.

# Kurzfassung

Um ein tiefgreifendes Verständnis unseres Planeten zu erlangen, untersuchen die Geowissenschaften die vielfältigen Prozesse des Systems Erde. Hierzu wird eine große Menge an Beobachtungs- und Simulationsdaten erfasst und analysiert. In diesen Daten manifestieren sich Prozesses als raum-zeitliche Muster. Um Erkenntnisse über die verschiedenen Prozesse des Systems Erde zu gewinnen, müssen diese Muster detektiert und interpretiert werden. Dies erfolgt durch Untersuchung der räumlichen und zeitlichen Variabilität in den Daten. Dies ist jedoch mit einigen Herausforderungen verbunden: (a) große raum-zeitliche Datenräume müssen betrachtet werden, (b) Aggregation und Dimensionreduktion sollte gering sein, um Informationsverlust zu vermeiden, (c) wichtige raum-zeitliche Muster müssen detektiert werden und (d) unter diesen Mustern müssen diejenigen identifiziert werden, die mit Prozessen des Systems Erde in Zusammenhang stehen. Die Dissertation untersucht, inwiefern Visual Analytics, d.h. die Kombination aus automatisierter Datenanalyse und interaktiver visueller Exploration, diese Herausforderungen addressieren und damit die Analyse von Prozessen in geowissenschaftlichen raum-zeitlichen Daten unterstützen kann. Hierzu fokussiert diese Dissertation auf drei wichtige Analyseperspektiven und addressiert jede einzelne anhand einer separaten Forschungsfrage: (1) Wie kann Visual Analytics die Detektion und Bewertung von wichtigen räumlichen Zuständen in den Daten unterstützen? (2) Wie kann Visual Analytics die Detektion und Bewertung von wichtigem zeitlichen Verhalten unterstützen? (3) Wie kann Visual Analytics die Analyse von zeitlichen Zusammenhängen in den Daten verbessern? Um jede dieser Forschungsfragen zu untersuchen, wurde jeweils ein neuartiger Visual Analytics Ansatz entwickelt. Der erste Ansatz erlaubt es, wichtige räumliche Zustände in den Daten sowie deren auftreten in der Zeit zu untersuchen. Hierzu werden *interactive visual summaries* eingeführt. Mittels hierarchischem Clustering werden alle in den Daten enthaltenen räumlichen Zustände in einer Clusterhierarchie verortet. Interaktive visuelle Analyse ermöglicht es, verschiedene räumliche Zustände aus den Daten zu extrahieren und die dazugehörigen raum-zeitlichen Muster zu interpretieren und zu bewerten. Der zweite Visual Analytics Ansatz unterstützt die Analyse des in den Daten zu beobachtenden zeitlichen Verhaltens sowie dessen Auftreten im geographischen Raum. Cluster Ensembles und interaktive visuelle Exploration ermöglichen die systematische Detektion und Interpretation verschiedener Typen zeitlichen Verhaltens. Der dritte Ansatz gestattet die Detektion und Analyse von zeitlichen Zusammenhängen in den Daten. Hierzu wurde eine etablierte Methode zur Analyse von zeitlichen Zusammenhängen zwischen zwei einzelnen Zeitreihen – gefensterte Kreuzkorrelation – durch Visual Analytics auf den Vergleich von Zeitreihenensembles erweitert. Durch die Erweiterung auf Ensembles ist es nicht nur möglich, Zusammenhänge zwischen Zeitreihen zu untersuchen, sondern auch Unsicherheiten in den Daten zu berücksichtigen. Die Visual Analytics Ansätze wurden anhand einer nutzer- und aufgabenorientierten Methodik entwickelt und erfolgreich in Anwendungsfällen aus der Erdsystem-Modellierung, der Ozeanmodellierung, der Paläoklimatologie und sogar den Kognitionswissenschaften eingesetzt. Die vorliegende Dissertation zeigt, dass die oben genannten Herausforderungen erfolgreich mit Visual Analytics addressiert werden können. Darüber hinaus wird deutlich, dass Visual Analytics aus folgenden Gründen einen wertvollen Ansatz zur Analyse von Prozess-bezogenen Mustern in raum-zeitlichen Daten darstellt: (1) geowissenschaftliche Anwendungen profitieren stark von der Interaktivität von Visual Analytics Ansätzen da diese eine intuitive Datenanalyse ermöglicht, (2) Visual Analytics kann die Grenzen existierender Analysemethoden erweitern und (3) es ermöglicht neue, aufschlussreiche Sichtweisen auf Daten und die darin beschriebenen Prozesse.

# Contents

# List of Figures

# Chapter I

# Introduction

# 1 Studying processes in geoscientific spatiotemporal data – Analysis perspectives and challenges

The geosciences and their manifold disciplines study the evolution and intricacies of our planet. To this end, researchers try to discover, analyze, and compare the different processes within the complex Earth system: from large-scale processes, such as crustal movement or global ocean and climate dynamics, to the smallest microbial and chemical processes in the soil. These studies help to identify and address important current and future challenges for mankind, such as global warming and its various consequences [65].

Processes of the Earth system are generally studied through analysis of spatiotemporal data. Technological advancement in the last decades has dramatically increased the amount of data available for such analysis. A plethora of observational and simulated data is provided by more and more powerful monitoring devices, computer hardware, and simulation models. Since a similar development can also be observed in many other scientific disciplines, the term *data-intensive science* was introduced to describe a new paradigm for scientific research [56]. This paradigm complements the existing empirical, theoretical, and simulation-based approaches, and focuses on data exploration to generate hypotheses and gain scientific insight into real-world phenomena. Data-intensive science requires an eScience infrastructure which supports capture, curation, and analysis of large amounts of data. Visual exploration is of particular importance in the data analysis part to extract information and meaning from the data [45, 56].

In the geosciences, it is information about processes, such as ocean currents, seasonal cycles, or deforestation, that has to be extracted from spatiotemporal data. These processes exhibit particular spatial and temporal behavior which manifests itself in the data as specific patterns. An example that shows that this behavior can be quite complex is the El Niño phenomenon [3]. It consists of a weakening of the equatorial trade winds off the South American coast, causing an increase in surface water temperature in the equatorial Pacific. This triggers a number of other meteorological and oceanographic phenomena across the globe, such as strong tropical storms [149]. El Niño occurs irregularly every two to seven years and lasts between nine months and two years. Since it is a large-scale phenomenon of significant climatological and economic impact, it has been studied intensely. The spatiotemporal patterns of previous El Niño events are well known and can be found in, e.g., observational data of the global ocean. Apart from El Niño, however, the Earth system comprises a multitude of other processes; many of these are not as well understood as the prominent El Niño

example.

In order to gain a better understanding of the various processes within the Earth system, scientists must capture their complex behavior by considering the geospatial and temporal variability in the data. Several analysis perspectives can be adopted to address this task. This thesis focuses on three perspectives:

**Perspective 1:** ***What prominent types of spatial situations can be found in the data and when do they occur in time?*** One way of approaching spatiotemporal data is to focus on the distribution of numerical values in geographic space – the spatial situation – and how it changes over time. Detecting the prominent types of spatial situations and their associated time periods provides a compact overview of the spatiotemporal dynamics in the data, including, e.g., seasonal cycles.

**Perspective 2:** ***What prominent types of temporal behavior can be found in the data and where are they located in space?*** Some processes manifest themselves in the data as particular temporal behavior that is bound to specific geographic areas, e.g., the Antarctic Circumpolar Current [117], a strong ocean current flowing around Antarctica. Therefore, an alternative way of analyzing spatiotemporal data is to concentrate on the different types of temporal behavior in the data and see where in geographic space they can be observed.

**Perspective 3:** ***Are there interrelations of temporal behavior in the data and what do they look like?*** Processes of the Earth system are often related in terms of their temporal behavior. For example, an El Niño event in the equatorial Pacific typically causes a warming of the tropical Atlantic. This warming, however, occurs four to five months after the initial El Niño [42]. Hence, the temporal behavior at different geographic regions or of different processes is often interrelated. Detecting and studying such interrelations provides scientists with deeper insight into the mechanics of the Earth system.

Examination of geoscientific applications has shown that analyzing spatiotemporal data from these three analysis perspectives poses several challenges:

**Challenge A:** ***Taking the entire data space into account.*** Due to the complexity of processes and the volume of geoscientific data, scientists often do not consider the entire data space and focus only on subsets, e.g., particular geographic regions or time periods. As a consequence, the insight about processes gained from these subsets cannot necessarily be extrapolated to other geographic regions or time spans covered by the data.

**Challenge B:** ***Reducing the amount of aggregation in the analysis.*** The analysis of large spatiotemporal data typically involves significant aggregation. For example, to cope with a large number of georeferenced time series and to preserve their geospatial context, scientists often aggregate the time dimension. Although scientists perform the aggregation based on experience and a well-grounded body of expert knowledge, it results in loss of information. The chosen aggregation focuses the analysis on a particular characteristic of temporal behavior; information about other potentially important aspects of temporal behavior is lost.

**Challenge C:** ***Detecting the most prominent patterns.*** Scientists often apply automated analyses to extract patterns from large geoscientific spatiotemporal data. The outcome of such automated analyses depends on the applied algorithm and its parameterization. The results obtained with a particular algorithm and parameterization, however, can only reflect certain spatial or temporal aspects. They may fail to represent the patterns that are most important to the analysis task.

**Challenge D:** ***Interpreting detected patterns.*** Any analytical result needs to be interpreted by domain experts. Based on their domain knowledge, they must assess which of the detected patterns are related to processes within the Earth system, and study them to gain a better understanding of the associated processes. They also have to decide which parts of the data may contain additional information and, hence, need further analysis, and which parts of the data contain merely noise or systematic measurement errors. However, researchers often do not understand in detail all individual steps behind the complex algorithms applied in the automated data analysis. In addition, the static plots typically used to assess the results do not convey all information necessary for a comprehensive interpretation.

In order to advance the analysis of processes in geoscientific spatiotemporal data, these challenges must be met.

## 2 Visual analytics as a means to explore geoscientific spatiotemporal data

Visual analytics has emerged as a promising interdisciplinary approach to the analysis of massive and complex data [77, 133]. It is "the science of analytical reasoning facilitated by interactive visual interfaces" [133, p. 4]. Visual analytics aims at creating "software that maximizes human capacity

to perceive, understand, and reason about complex and dynamic data and situations" [133, p. 6]. To this end, it combines automated data analysis and interactive visualization to facilitate the analytical discourse with the data [77, 133]. The automated analysis part typically includes techniques from statistics or data mining [50] to discover patterns in data. The combination of statistics with visualization to enable exploratory data analysis alongside the traditional confirmatory analysis approach has already been advocated by John W. Tukey in the 1970s [140]. He also stressed the power of visualization in this process: "The greatest value of a picture is when it forces us to notice what we never expected to see" [140, p. vi]. This would later be substantiated by the information visualization community which builds upon computers and graphical user interfaces to develop effective techniques for interactive visual data analysis [24, 27, 129, 148]. Many visual analysis and visual data mining solutions [76] are guided by Ben Shneiderman's *visual information seeking mantra* "overview first, zoom and filter, then details-on-demand" [30, 127]. To support the analysis of increasingly large and complex data, visual analytics aims at an even tighter integration of automated and visual analysis [78]. Keim's *visual analytics mantra* illustrates this integration: "Analyze first, show the important, zoom, filter and analyze further, details on demand" [78, p. 7]. Visual analytics uses computers to algorithmically detect patterns in large and complex data, and interactive visualization to exploit human perception and cognition to recognize patterns, to assess them, and to put them into context. Allowing users to visually explore and interact with data in such a way enables domain experts to make full use of their domain knowledge and intuition in the analysis process and to obtain a detailed mental model of the data space and the phenomena it describes.

In the last decade, visual analysis and visual analytics have been successfully employed in various application areas [73], e.g., medicine, physical sciences, or engineering. The geovisual analytics community specifically focuses on the exploration of geographical data. Geovisual analytics approaches combine techniques from information visualization, cartography, geovisualization and geographical information systems. Andrienko and Andrienko [9] present a taxonomy of general tasks that are relevant in the exploration of geographical data, as well as examples for potential solutions. Geovisual analytics has been applied to, e.g., analysis of spatiotemporal variation of mobile phone usage [6], car and vessel movement data [10, 33], or for exploration and prediction of crimes in the United States [7, 97]. Many more examples of geovisual analytics solutions that facilitate visualization and analysis of spatiotemporal data are presented by Dykes et al. [39] and Dogde et al. [36].

Although these examples demonstrate that the combination of automated and visual analysis constitutes a particularly powerful approach to the exploration of spatiotemporal data, few tools explicitly

support geoscientific scenarios, such as studying the temporal evolution of glaciers in Greenland [38], validation of geoscientific simulation models of glacial isostatic adjustment [142] or paleoclimatic cold events [72], hypothesis generation about atmospheric climate change [74], or analyzing periodicity in ocean data [48]. This may be due to the complexity of geoscientific application problems and phenomena under study – which also applies to the physical sciences in general [90] – but is also somewhat surprising since visual analytics has great potential to facilitate the analysis of process-related patterns in geoscientific spatiotemporal data.

## 3 Hypothesis and research questions

The overall aim of this thesis is to study how visual analytics can facilitate the analysis of processes in geoscientific spatiotemporal data. It is based on the hypothesis that the challenges outlined in Section 1 can be addressed with visual analytics. Statistical and data mining methods [50] will enable geoscientists to take the entire spatiotemporal data space into account, instead of limiting themselves to particular subsets prior to the analysis (Challenge A). Popular techniques to extract information from large or complex data spaces include clustering, data reduction, or dimensionality reduction [73]. They algorithmically structure and aggregate the data by detecting prominent patterns in the data. Furthermore, data mining techniques that are able to cope with multidimensional data will reduce the amount of aggregation required at the early stages of the analysis process (Challenge B). The combination with interactive visual analysis will allow scientists to interpret the results of the automated analysis, to assess the detected patterns in their spatiotemporal context, and to obtain an overview of the most prominent patterns in the data (Challenge C). Effective visualization and interaction techniques facilitate the identification of patterns that are related to environmental or geophysical processes, as well as detailed inspection of these patterns. In particular, interactive visual interfaces allow scientists to focus on specific patterns, gain a detailed understanding of their spatial, temporal, and statistical characteristics, and use their domain knowledge, experience, and intuition in the interpretation process [77, 78, 133] (Challenge D).

In order to facilitate the analysis of processes in geoscientific spatiotemporal data, each of the three analysis perspectives presented in Section 1 has to be supported. Consequently, this thesis investigates each of the following three research questions in a separate chapter:

**Research question 1 (Chapter III): How can visual analytics support the detection and assessment of prominent types of spatial situations?** This chapter investigates how automated analysis can be coupled with mechanisms that support assessment and refinement of the analytical results to facilitate detection of the most prominent spatiotemporal patterns. For this purpose, this chapter studies how clustering can be integrated with interactive visual analysis to enable geoscientists to study the spatial and temporal variability in the data, and to create a compact visual representation of the most prominent types of spatial situations and their occurrence over time. This chapter further describes how the proposed approach is applied to detect processes in ocean data.

**Research question 2 (Chapter IV): How can visual analytics support the detection and assessment of prominent types of temporal behavior?** To address this question, this chapter examines how the comparison of a large number of georeferenced time series can be supported while simultaneously reducing the amount of data aggregation and dimensionality reduction required. To this end, it investigates how cluster ensembles [131] can be used to detect prominent types of temporal behavior, and how interactive visual exploration allows scientists to analyze and compare the different patterns and to relate them to processes within the Earth system. The proposed visual analytics approach is developed in an ocean modeling context and applied to the assessment of simulation model output.

**Research question 3 (Chapter V): How can visual analytics improve the analysis of interrelations of temporal behavior?** For a thorough detection of interrelations of temporal behavior in geoscientific data, many geoscientific applications, such as climate modeling [21] or paleoclimatology, require the comparison of ensembles of time series. This chapter examines how an established technique for comparison of two individual time series, windowed cross-correlation [4, 20], can be enhanced by visual analytics to facilitate the detection of interrelations between entire ensembles of time series. The benefits of the proposed visual analytics approach are studied in use cases from paleoclimatology and cognitive science.

## 4 Structure of this thesis

This thesis comprises six chapters. After the introduction, Chapter II describes the exemplary use cases in which the outlined research questions were studied. It also provides details about the user- and task-centered methodology that underlies the research presented in this thesis. The core of this

thesis is formed by Chapters III through V, each addressing one of the identified research questions. Each of these three chapters were written as stand-alone articles and published in peer-reviewed journals as follows:

**Chapter III:** P. Köthur, M. Sips, A. Unger, J. Kuhlmann, and D. Dransch. Interactive Visual Summaries for Detection and Assessment of Spatiotemporal Patterns in Geospatial Time Series. *Information Visualization*, 13(3):283–298, doi:10.1177/1473871613481692, 2014.

**Chapter IV:** P. Köthur, M. Sips, H. Dobslaw, and D. Dransch. Visual Analytics for Comparison of Ocean Model Output with Reference Data: Detecting and Analyzing Geophysical Processes Using Clustering Ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1893–1902, doi:10.1109/TVCG.2014.2346751, 2014.

**Chapter V:** P. Köthur, C. Witt, M. Sips, N. Marwan, S. Schinkel, and D. Dransch. Visual Analytics for Correlation-Based Comparison of Time Series Ensembles. *Computer Graphics Forum*, 34(3):411–420, doi:10.1111/cgf.12653, 2015.

This thesis concludes with a synthesis of the presented results (Chapter VI), which summarizes the answers to the investigated research questions, presents the main conclusions of this thesis, and briefly outlines opportunities for future research.

# Chapter II

# Methodology

# 1 Use cases

Due to the application-oriented nature of visual analytics, developing effective solutions requires close collaboration with domain experts. Hence, the three research questions outlined in Chapter I, Section 3 were studied in three exemplary use cases. In each use case, domain experts from the German Research Centre for Geosciences GFZ[1] and the Potsdam Institute for Climate Impact Research[2] adopted a particular analysis perspectives (Chapter I, Section 1) in the context of a specific geoscientific application.

The first use case addresses the detection of prominent types of spatial situations in ocean data. The objective of the second use case is to detect prominent types of temporal behavior and to utilize these finding in the assessment of ocean simulation models. Both use cases focus on univariate, scalar data on a regularly structured two-dimensional grid. This type of gridded data is not only studied in oceanography but also in other disciplines such as climatology or landscape ecology. There are two important ways of conceptualizing such data [7]. Both perspectives represent complementary views on the spatial and temporal variability in geoscientific data. The first use case considers the data as an ordered sequence of two-dimensional geospatial distributions of scalar values (spatial situations). The data may contain from a few hundreds to several thousands of spatial situations, one for each time stamp. The second use case considers the data as a set of geographically referenced temporal profiles. The grid points of the data grid represent geographic coordinates and are associated with regularly sampled temporal profiles (or time series), describing the temporal behavior at these coordinates. The data may easily comprise thousands or hundreds of thousands of grid points and, therefore, just as many temporal profiles. The third use case aims at supporting paleoclimatologists and time series analysis experts in the detection and assessment of interrelations between uncertain time series derived from stalagmites. To this end, ensembles of time series are compared. These time series are regularly sampled and similar to the temporal profiles associated with the grid points in the gridded data.

In each use case, the collaborating scientists are working in a data-intensive scientific environment where analysis of spatiotemporal data is an integral part of their daily research routine. Not only do they analyze data to confirm hypotheses about processes but they also try to generate new hypotheses from the observed patterns. The data are then studied with respect to these hypotheses to gain new insights into the processes under study.

[1] http://www.gfz-potsdam.de
[2] http://www.pik-potsdam.de

# 2 User- and task-centered development

For visual analysis and visual analytics solutions to be effective, they have to focus on the needs of the users [77]. This requires a thorough understanding of how users approach the analysis of their data. An effective strategy to describe this process is to identify the individual analysis tasks that users perform [8,9,122]. To this end, various task taxonomies have been proposed in the last decades. The taxonomies by Peuquet [111] or Blok [19], for example, focus on the characteristics of spatiotemporal data to define high-level analysis tasks. Andrienko and Andrienko's [9] task taxonomy is also data-oriented, but more specific and explicitly supports exploration of spatiotemporal data. In contrast, Amar and Stasko [5] focus on the user and define high-level *knowledge tasks* that have to be addressed in complex analyses. Shneiderman's popular *task-by-data-type taxonomy* [127] describes a general, and also very high-level, approach to visual exploration. Wehrend and Lewis [150], Casner [25], or Zhou and Feiner [158], on the other hand, focus on rather low-level perceptual and cognitive tasks that users have to solve when working with visualizations. Knapp [79] presents a practical guide to task analysis and visualization design. The approach is based on a thorough analysis of the user's reasoning process and associated task structure, development of a task model and, eventually, a design model that suggests specific visualizations.

In previous work, the author introduced a user- and task-centered methodology which extends the approaches of Knapp as well as Wehrend and Lewis, and combines it with several knowledge elicitation techniques [37]. The visual analytics solutions presented in this thesis were developed following this rigorous approach. It comprises three main steps: task analysis, visual analytics concept and implementation, and evaluation.

## 2.1 Task analysis

First, the geoscientists' reasoning process in the analysis of spatiotemporal data was examined to derive a *domain task model*. The model consists of domain-specific descriptions of the analysis tasks that have to be accomplished. The first version of such a model was obtained through an initial unstructured interview in which the collaborating geoscientists were asked to describe how they approach the analysis. The interview provided first insights into the specific application problem, the involved challenges, the domain experts' reasoning process, and the requirements and analysis tasks that had to be supported. Subsequent semi-structured interviews were used to acquire additional information where necessary, and to discuss and refine the domain task model. In the second stage of the task

analysis, *cognitive actions* were determined for every identified domain task. These actions describe the required visual interaction with the display, e.g., *identify*, *locate*, *compare*, *cluster*, or *associate*. The verbal description of the analysis process obtained through think-aloud techniques in the preceding interviews was of great help at this stage. Quite often a cognitive action is relevant to several domain tasks. Whenever possible, they were summarized into classes of cognitive actions, yielding a *cognitive actions model*.

## 2.2 Visual analytics concept and implementation

The cognitive actions model is the basis for developing the visual analytics concept. The spatial, temporal, or statistical information which has to be explored in the visual analytics solution is determined by the respective cognitive action. In order to support the cognitive actions and associated domain tasks, suitable automated analysis methods as well as visualization and interaction techniques had to be identified. First, several candidate algorithms were chosen from the various data mining and statistics approaches available. These algorithms were evaluated regarding their ability to detect patterns in geoscientific spatiotemporal data and their potential to support the respective cognitive actions. This included testing the algorithms on well-understood real-world data. Furthermore, various similarity measures for geoscientific spatiotemporal data were tested and potential performance bottlenecks identified and assessed. Based on the findings of several weeks to months of experimentation, the algorithms for the automated analysis were chosen. In a second step, interactive *visualization components* were developed to enable exploration and interpretation of the automated analysis results. These tailored components typically comprise multiple linked views and include techniques from cartography [18, 128], geovisualization [39], or information visualization [129, 148]. Whenever established visualization methods and interaction mechanisms did not suffice to support a particular cognitive action, novel techniques were developed and integrated with the automated analysis part. Finally, prototypical implementations of the derived visual analytics concepts were evaluated.

## 2.3 Evaluation

User-based, result-oriented, formative evaluation was applied to every visual analytics solution. In this iterative process, the collaborating scientists evaluated the prototypes in practice with regard to their ability to facilitate their tasks and reasoning process. The advantages and disadvantages of individual visualization components were discussed in informal interviews but also in spontaneous, informal meetings and discussions. The results of these discussions were considered in the next

iteration of the approach.

# Chapter III

# Interactive visual summaries for detection and assessment of spatiotemporal patterns in geospatial time series

**Published as:** P. Köthur, M. Sips, A. Unger, J. Kuhlmann, and D. Dransch. Interactive Visual Summaries for Detection and Assessment of Spatiotemporal Patterns in Geospatial Time Series. *Information Visualization*, 13(3):283–298, doi:10.1177/1473871613481692, 2014.

**Abstract.** Numerous measurement devices and computer simulations produce geospatial time series that describe a wide variety of processes of system Earth. A major challenge in the analysis of such data is the complexity of the described processes, which requires a simultaneous assessment of the data's spatial and temporal variability. To address this task, geoscientists often use automated analyses to compute a compact description of the data, ideally comprising characteristic spatial states of the process under study and their occurrence over time. The results of such automated methods depend on the parameterization, especially the number of extracted spatial states. A particular number of spatial states, however, may only reflect certain spatial or temporal aspects. We introduce a visual analytics approach that overcomes this limitation by allowing users to extract and explore various sets of spatial states to detect characteristic spatiotemporal patterns. To this end, we use the results of hierarchical clustering as a starting point. It groups all time steps of a geospatial time series into a hierarchy of clusters. Users can interactively explore this hierarchy to derive various sets of spatial states. To facilitate detailed inspection of these sets, we employ the concept of interactive visual summaries. A visual summary is the depiction of a set of spatial states and their associated time steps or intervals. It includes interactive means that allow users to assess how well the depicted patterns characterize the original data. Our visual interface comprises a system of visualization components to

facilitate both the extraction of sets of spatial states from the hierarchical clustering output and their detailed inspection using interactive visual summaries. This study results from a close collaboration with geoscientists. In an exemplary analysis of observational ocean data, we show how our approach can help geoscientists gain a better understanding of geospatial time series.

# 1 Introduction

Geospatial time series describe a broad range of processes of System Earth, such as atmospheric circulation, animal migration, or river runoff, just to name a few. Typically, these data either stem from measurement devices, for example, satellite sensors, tide gauges, or global positioning system (GPS) sensors, or from computer simulations, such as environmental simulation models. In numerous applications such as risk assessment or civil engineering, it is crucial to understand the processes described by these data. Gaining this understanding is a challenging task because scientists need to assess the data's spatial and temporal variability simultaneously.

In this article, we focus on time series where each time step represents a regular two-dimensional (2D) spatial distribution of scalar values, which we call a spatial situation. An important objective is to find spatiotemporal patterns that capture the data's variability. To this end, scientists often apply automated analyses to, first, extract the characteristic spatial situations and, second, to assign the individual time steps to these situations. The result of this analytical procedure – a limited number of characteristic spatial situations and their occurrence over time – is used as a compact description of the time series. It allows scientists to assess the data's spatial and temporal variability by focusing on the most important spatiotemporal information.

The outcome of such automated analyses depends on the applied algorithm and its parameterization. An important parameter is the number of spatial situations to extract from the data. A particular number of spatial situations, however, may only reflect certain spatial or temporal aspects of a geospatial time series. Hence, any analytical result should only be regarded as a proposal of characteristic spatial situations, a proposal that needs assessment by domain experts. They must be given the means to decide, based on their expert knowledge about real-world processes, whether the analytical result contains patterns that are important to the analysis task and which aspects need further refinement.

In this article, we introduce a visual analytics approach that allows scientists to extract and explore various sets of spatial situations to detect characteristic spatiotemporal patterns in the data. For this purpose, our approach uses hierarchical clustering to aggregate all spatial situations in a time series into a hierarchy of clusters; a cluster is a set of similar spatial situations. For each cluster, we compute a representative spatial situation. Users can interactively explore this hierarchy to derive different sets of representative spatial situations. To facilitate detailed assessment of these sets, we employ the concept of interactive visual summaries. A visual summary is the depiction of a set of spatial

situations and their associated time steps or intervals. It includes interactive means that allow users to assess how well the depicted patterns characterize the original data. This approach results from close collaboration with geoscientists and a thorough task and requirement analysis.

We present a visual interface that facilitates both the extraction of sets of spatial situations from the hierarchical clustering output and their detailed inspection using interactive visual summaries. Users can interactively select clusters from the hierarchy and assess the corresponding spatiotemporal patterns in a visual summary. Exploiting the hierarchical structure of the clustering output, users can interactively split or merge any cluster in a visual summary and easily assess the resulting changes. We provide visualization components that let users decide whether a particular pattern in a visual summary represents important information, and which clusters should be refined via split or merge operations. This exploration process enables scientists to detect spatiotemporal patterns that they consider characteristic.

In particular, the contributions of this article are as follows:

1. We present a design study that is the result of a close collaboration with users and a thorough task and requirement analysis.

2. We introduce an analytical approach that allows users to explore and summarize a geospatial time series by extracting and refining different sets of spatial situations from the data.

3. We show that the exploration of visual summaries enables users to identify spatial situations that they consider characteristic.

4. We show that the exploration of visual summaries can lead to a better understanding of geospatial time series.

The rest of this article is organized as follows: After reviewing related work, we provide an overview of our concept and explain the design requirements for a visual exploration tool that facilitates extraction and exploration of spatiotemporal patterns in geospatial time series. Next, we provide a detailed description of the applied clustering algorithm, as well as the visual exploration tool and its visual encoding. We further demonstrate and discuss the utility of our approach in an exemplary analysis of observational ocean data. We conclude with a short summary and with potential future research directions.

# 2 Related work

In this section, we briefly discuss examples from the geosciences for the automated detection of spatiotemporal patterns and explain why we chose hierarchical clustering. We further review studies on interactive visualization for spatiotemporal data analysis because visualization has proven to be valuable for incorporating users' domain knowledge and gaining insight into time series [2]. Since we use hierarchical clustering as an automated analysis step, we also discuss approaches that facilitate interactive visual exploration of clustering parameters and cluster hierarchies.

## 2.1 Analytical methods for detecting spatiotemporal patterns – Examples from the geosciences

Geoscientists apply, and often combine, various computational methods to detect prominent spatiotemporal patterns in environmental time series. FEM-K-trends [60] and T-mode principal component analysis [64] transform or reduce the number of dimensions of geospatial time series. The basic assumption is that a limited number of principal components express enough of the spatiotemporal structure of the data. Gaussian mixture models and expectation maximization [119] as well as clustering algorithms such as k-means or hierarchical clustering [63, 69] sort observations into k groups such that the similarity is high among members of the same group and low between members of different groups. The identified clusters provide a condensed description of the original data (for further readings, please refer to studies by Jain et al. [67], Miller and Han [102], or Han and Kamber [50]). Another popular clustering and dimensionality reduction technique in the geosciences is self-organizing maps [55, 61].

A substantial problem with all these methods is the parameterization of the algorithms. A good parameterization should result in a few spatial situations that represent characteristic states of the process under study. An important parameter is the number of clusters. Choosing the number of clusters is a conceptual difficulty in clustering. This parameter is often specified a priori by users or determined with the help of statistical measures, such as the silhouette coefficient [118] or the Bayesian information criterion [123].

We chose agglomerative hierarchical clustering as the computational method for the automated analysis because it does not require specifying the number of clusters in advance. In our approach, the hierarchy of clusters is a starting point for exploring different sets of spatial situations extracted from the data. In a recent publication [81], we demonstrated that hierarchical clustering can capture

characteristic patterns in geospatial time series.

## 2.2 Interactive visualization for spatial time series analysis

Established techniques for visualizing spatiotemporal data are small multiples and map animation [11, 138, 139]. While these techniques are effective for small time series, they do not scale well to larger time series due to limited screen space or perceptual and cognitive limitations such as change blindness [44, 141].

Interactive visualization allows for analyzing large time series by facilitating the information-seeking mantra: "Overview first, zoom and filter, then details on demand" [127]. Typically, one or several overview visualizations present the data in aggregated form, while multiple coordinated views allow users to formulate queries against the data and assess the results in detail. Clustering is a common means of creating a compact description of data and can serve as a starting point for interactive exploration. Many successful approaches combine clustering and interactive visualization to facilitate analysis of, for example, nonspatial time series data [51, 89, 146] or time-varying vector fields [145, 153].

Approaches focusing on geospatial time series are less numerous. Bruckner and Möller [23] use density-based clustering to interactively explore spatiotemporal data. Their approach is tailored to visual effects design, a different application problem requiring other visualization and interaction techniques. Frey et al. [48] extract similarity lines from similarity matrices to assess and compare temporal behavior in (geo)spatial time series. They focus on the detection of recurring patterns, while we allow users to detect various types of characteristic patterns. More closely related to our research is the study by Andrienko et al. [7]. The authors use self-organizing maps to cluster the spatial situations of a geospatial time series and link the results with interactive displays that visualize the extracted spatial patterns and their occurrence over time. Their concept facilitates exploration of spatiotemporal pat- terns in a single clustering result, that is, a single partitioning of the data into clusters. In contrast, our goal is to support exploration and assessment of many different partitionings of the data. We seek to help scientists arrive at an appropriate partitioning of the data into clusters that captures those patterns that they consider characteristic and important to the analysis task. This requires a different clustering approach as well as a different visualization and interaction concept.

### 2.3 Interactive exploration of cluster hierarchies

Depicting a hierarchy of clusters (dendrogram) is essentially a tree visualization problem. Herman et al. [54] provide a comprehensive survey of typical application areas and key issues from an information visualization perspective.

Many studies facilitate exploration of hierarchical clustering results. The Hierarchical Clustering Explorer [124, 125] integrates a dendrogram with color mosaics and 2D scattergrams for analyzing genomic microarray data. Kreuseler and Schumann [83] introduce an algorithm for computing an abstraction of a dendrogram. They also propose Magic Eye View as a focus & context technique to map the resulting hierarchy graph onto the surface of a hemisphere. Chen et al. [28] combine an abstract overview dendrogram with detail- view dendrograms and reorderable matrices to facilitate exploration of multivariate data. SpectraMiner [156] combines an interactive radial dendrogram with other linked views to analyze high-dimensional, nonspatial data. This approach is later extended in the ClusterSculptor [106] system to allow for interactive refinement of cluster hierarchies. MultiClusterTree [144] visualizes a cluster hierarchy in a 2D radial layout and combines it with circular parallel coordinates and other views.

The described techniques address nonspatial and/or nontemporal data. Analyzing hierarchical clustering results for geospatial time series, however, requires a combined assessment of the data's spatial and temporal domain. To facilitate this combined assessment, our visualization design integrates techniques from geovisualization, time series visualization, and graph visualization. We use the dendrogram to let users derive different sets of spatial situations from the data, but the primary focus is on exploring and visualizing the spatiotemporal information in the corresponding visual summaries.

## 3 Concept and design requirements

In this section, we present a visual analytics concept that is the result of a close collaboration with Earth system modelers, hydrologists, and ocean modelers. We adopted a user-centered and task-centered approach [37] to derive a thorough understanding of the challenges that scientists face when they are studying geospatial time series. From the findings of our analyses, we derived the following twofold concept:

1. *Hierarchical clustering.*

   We use agglomerative hierarchical clustering to construct a binary tree that aggregates the spatial situations associated with individual time steps into a hierarchy of clusters. Users do not have to specify the number of clusters in advance but rather use the hierarchy as a means to extract and explore spatial situations from the data.

2. *Interactive exploration.*

   A visual exploration tool allows users to traverse the dendrogram from top to bottom to progressively extract varying sets of spatial situations from the geospatial time series and explore the associated spatiotemporal patterns using visual summaries. Since the dendrogram represents a hierarchy of clusters, a top-down traversal enables users to assess spatiotemporal patterns at different levels of detail. The interactive visual summaries allow users to identify clusters that represent characteristic spatiotemporal patterns.

We further identified the following design requirements (DRs) for a visual exploration tool that facilitates interactive extraction and exploration of spatiotemporal patterns in geospatial time series:

DR1 *For a specific selection of clusters from the dendrogram, present the corresponding visual summary to users.*

To assess the spatiotemporal context of patterns, scientists must know what extracted spatial situations look like and when these situations occur in the time series.

DR2 *Enable users to gradually increase the level of detail of spatiotemporal patterns presented to them.*

Scientists do not have a complete understanding about which patterns are hidden in geospatial time series. Therefore, they prefer to gradually explore the spatial situations and their occurrence over time, starting with a rather coarse visual summary, and refining this summary in a stepwise manner.

DR3 *Provide information about the level of detail.*

To give scientists orientation in the exploration process, they need to be aware of the current degree of refinement.

DR4 *Allow users to assess the quality of a visual summary.*

Users want to know how well the clusters represent the original time series data and refine clusters where necessary.

DR5 *Allow users to visually detect periodic or quasi- periodic patterns.*

Recurring patterns can be an important aspect of geospatial time series. Scientists want to be able to visually detect and assess such recurrences in a visual summary.

In the following, we will describe the hierarchical clustering algorithm and our visual exploration tool.

# 4 Hierarchical clustering

In this section, we describe our experience with applying hierarchical clustering to geospatial time series, explain our choice of the linkage method, and discuss feedback from geoscientists.

## 4.1 Algorithm

Hierarchical clustering groups data objects into a tree of clusters. This grouping can be performed either by iteratively dividing the set of data objects or by agglomerating the data objects. In our approach, we apply agglomerative hierarchical clustering. It considers each item of a data set a single cluster. In each iteration of the clustering process, the two clusters $p$ and $q$ with the highest similarity are agglomerated into a new cluster $h = p \cup q$. The clustering process terminates if there is only one cluster left containing the entire data set. The input to agglomerative hierarchical clustering is a list of $[n, 2]$ dissimilarities of $n$ data items. The output is a binary tree representing the cluster hierarchy (see also Müllner [104] for a recent survey on agglomerative hierarchical clustering). The structure of the resulting cluster hierarchy depends strongly on the measure of dissimilarity and the agglomeration method.

In our scenario, the dissimilarity between time steps is based on the dissimilarity of their associated spatial situations. We conducted several experiments to assess different dissimilarity measures and obtained the best results with the sum of squared errors [91]. Let $i$ and $j$ be two time steps, and let $D_i$ and $D_j$ be their associated two-dimensional distributions of scalar values. Without loss of generality, we consider $D_i$ and $D_j$ as $N \times M$ matrices, and compute the dissimilarity $d(i, j)$ between time steps $i$ and $j$ with $d(i, j) = \sum_{k=1}^{N} \sum_{l=1}^{M} (D_i[k, l] - D_j[k, l])^2$. The computational effort to construct the list of dissimilarities depends on the resolution of the spatial situations and the number of time steps.

In collaboration with geoscientists, we tested various agglomeration methods with respect to their applicability to geospatial time series. The test data sets differed in terms of spatial and temporal

resolution, phenomenon described, and geographic area examined. On one hand, we applied the Lance and Williams [86] sorting strategy to these data to realize single linkage, complete linkage, average linkage, and minimum variance agglomeration. As an alternative to Lance–Williams, we used the Chameleon algorithm [70] as a representative for graph-based agglomeration strategies. It includes a preclustering similarity search that constructs a k-nearest-neighbor graph and partitions this graph into initial clusters. The agglomeration stage of this algorithm is based on the connectedness of cluster members and the proximity of clusters within the k-nearest-neighbor graph.

Based on the assessment by domain experts, we decided to use the average linkage method. We implemented the nearest neighbor chain (NN-chain) algorithm [105] because it is time-optimal for average linkage agglomeration. To find the closest pair of clusters, the algorithm constructs an NN-chain. Starting with an arbitrary time step $i$, an NN-chain is the sequence $\mathrm{NN}(i) = j, \mathrm{NN}(j) = k, \cdots, \mathrm{NN}(q) = p, \mathrm{NN}(p) = q$ where $j$ has the smallest dissimilarity to $i$ among all time steps, according to the dissimilarity between their associated spatial situations $D_i$ and $D_j$. The intercluster dissimilarity within an NN-chain decreases in a monotonic manner. A closest pair $p$ and $q$ of clusters is detected if the NN-chain arrives at a situation where $\mathrm{NN}(p) = q$ and $\mathrm{NN}(q) = p$. To determine a new closest pair of clusters, the algorithm computes a new NN-chain from the cluster that preceded $p$ and $q$, or from an arbitrary cluster if the NN-chain is empty. The time and space complexity of the algorithm for average linkage agglomeration is $O(n^2)$.

## 4.2 Discussion

Our collaborators clearly favored the average linkage method because it yielded easily interpretable results and was able to capture characteristic patterns in geospatial time series; the output of the other methods was often nonintuitive and difficult to interpret. In their daily work, however, they noted that clusters that they consider as similar are sometimes placed in distant parts of the hierarchy. After discussing this issue with the users, we identified two potential reasons. When domain experts visually compare two clusters, they sometimes do not consider the entire geographic and data domain, but focus on specific geographic regions and/or data value ranges. Since our measure of dissimilarity considers the entire geographic space and data domain, the computed dissimilarity may differ from expert judgment in such cases. We also noted that the experts' criteria for comparing any two clusters may change during the exploration process. Therefore, we cannot adjust the measure of dissimilarity to these criteria a priori. Another potential reason could lie in the strict nature of merge decisions during the agglomeration phase. However, our experiments with more flexible multiphase

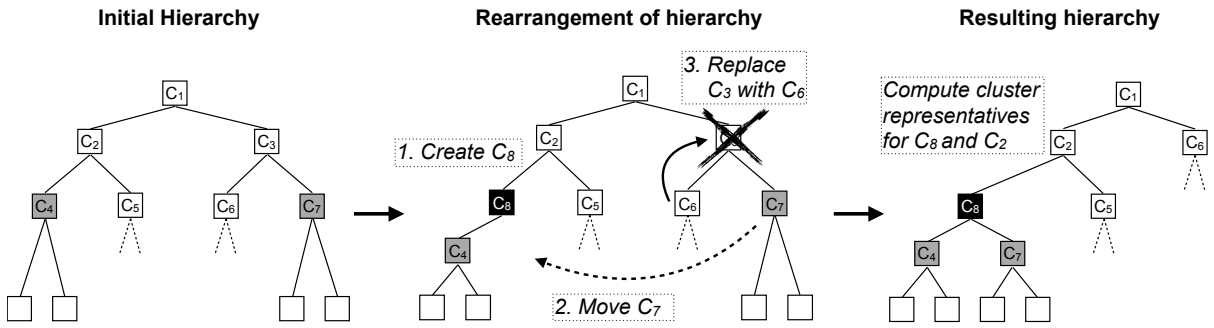



**Figure III.1:** Simplified example of a rebalancing of the cluster hierarchy. If users choose to merge clusters $C_4$ and $C_7$, a new agglomerative cluster $C_8$ – with $C_4$ and $C_7$ as its children – will be created to replace $C_4$. Meanwhile, cluster $C_3$ is removed and replaced by $C_6$.

hierarchical clustering methods, such as Chameleon, have not led to better results. Our experience suggests that the reason for this lies in the complexity of the data. Geospatial time series may describe many different, often nonlinear processes, singular events, or recurring patterns. Each parameterization of the clustering introduces additional assumptions about the data and often emphasizes only a particular aspect.

To address this issue, our visual interface allows users to interactively rearrange clusters in the dendrogram. This provides domain experts with the flexibility to bring the depicted cluster structure in the hierarchy in accordance with their expert judgment. The rebalancing of the cluster hierarchy can be achieved using standard tree sorting algorithms [101]. Figure III.1 illustrates this process and the resulting changes to the dendrogram.

## 5 Interactive exploration

We propose five tightly coupled, interactive visualization components – *hierarchy explorer*, *sequence view*, *sequence explorer*, *periodicity explorer*, and *spatial gallery*. The specific coupling of the visual components facilitates extraction of various sets of spatial situations from the data and detailed inspection of the corresponding spatiotemporal patterns using interactive visual summaries. In this section, we will first give an overview of the components (Figure III.2) and their visual and interactive coupling, before explaining the visual encoding in more detail.

### 5.1 Overview

The *hierarchy explorer* allows users to extract different sets of spatial situations from the data via drill-down, roll-up, and rearrangement operations on the cluster hierarchy. It depicts the representative

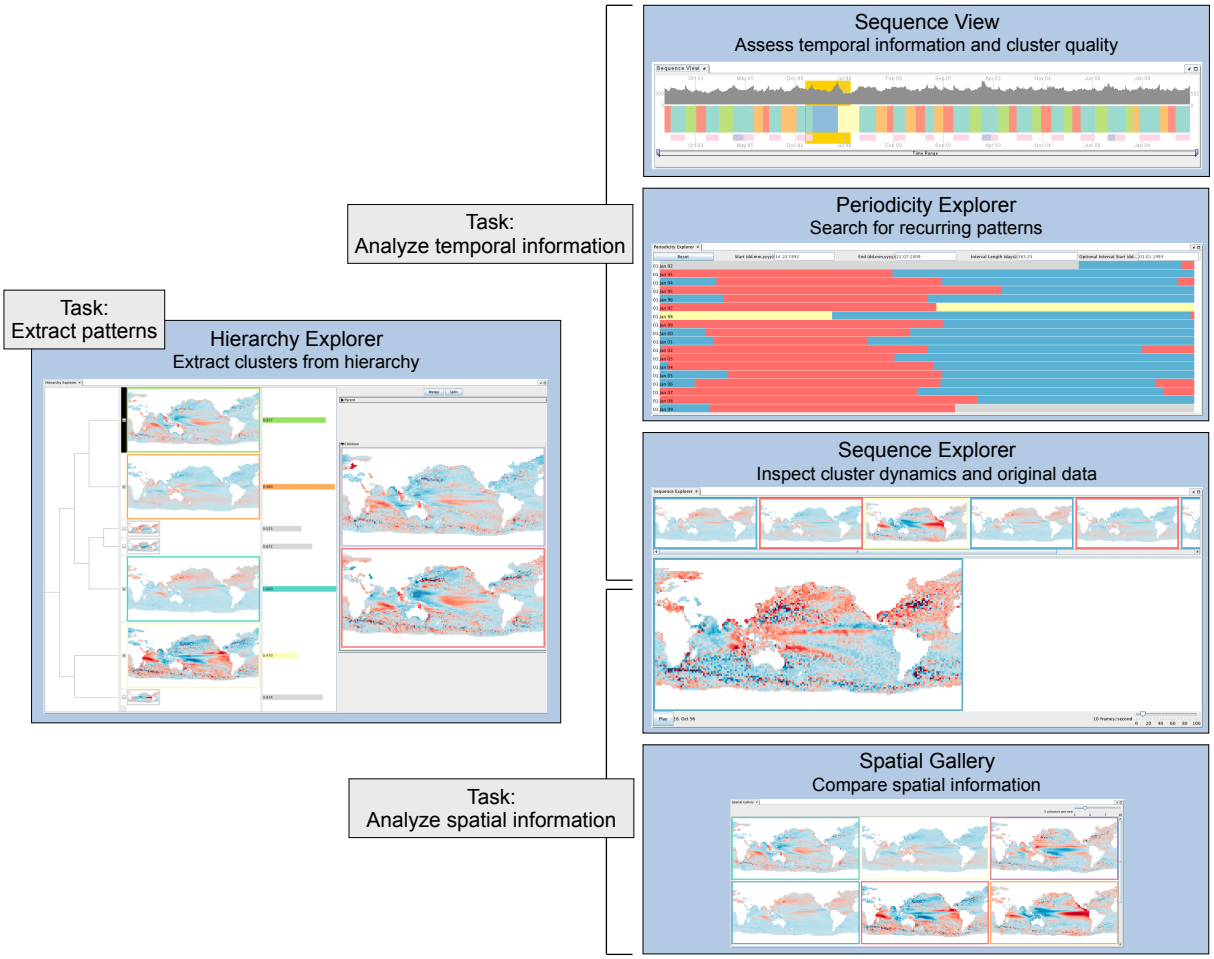



**Figure III.2:** Overview of the five visualization components that are integrated into our exploration tool and the supported tasks. The components are visually linked through a consistent color coding of the cluster affiliation of time steps and the cluster representatives.

spatial situations of the clusters (DR1) and indicates their position in the dendrogram (DR3). Users can select individual clusters to inspect the representative spatial situations of its two child clusters and decide whether they would like to split the selected cluster, merge it with its sibling, or focus on a different cluster (DR2). In addition, users can merge any two clusters and, thus, manually rearrange the hierarchy. It also provides statistical information that gives hints on the overall quality of the cluster representatives (DR4). The hierarchy explorer allows users to focus on specific patterns of interest in the current visual summary by selecting a subset of the extracted clusters. This subset will be forwarded to the spatial gallery, the periodicity explorer, and the sequence view.

The *sequence view* concisely depicts the temporal information of a visual summary. It shows at which time steps the currently extracted spatial situations occur (DR1) and provides hints on how well they represent these time steps (DR4). When users select a specific cluster, the sequence view

presents a preview of the potential changes in the cluster affiliation of time steps that would result from splitting this particular cluster (DR2). Additionally, users can select any time step or time period in the sequence view for further inspection in the sequence explorer.

The *sequence explorer* facilitates a more detailed assessment of spatiotemporal dynamics in the data. Users can inspect the temporal order of clusters in a visual summary (DR1) and compare the spatial situations associated with each time step with their respective cluster representative (DR4). The periodicity explorer focuses on recurrences in a visual summary by supporting visual detection of (quasi-)periodic patterns (DR5).

The *spatial gallery*, in combination with the sequence view, is an effective mechanism of presenting the (intermediate) results of the exploration process (DR1). It exclusively focuses on the cluster representatives in the current visual summary, summarizing the spatial information and facilitating comparison of spatial patterns. To establish a visual link between all five views, we use a color scheme that is consistent across all five visualization components to encode cluster affiliation of time steps and representative spatial situations.

We provide a flexible framework that allows free arrangement of the five visualization components. Depending on the available screen space or number of displays, users can choose to arrange the views in single windows, a flexible matrix layout, tabs, or any combination of these modes.

## 5.2 Visual encoding

### Hierarchy explorer

The hierarchy explorer consists of two main components (Figure III.3): an overview dendrogram, presenting the hierarchy of clusters at a user-specified level of detail (DR3) as well as providing information about the cluster quality (DR4), and a spatial preview, allowing users to preview the spatial information in the child clusters (DR1). Both components facilitate merge or split operations in the cluster hierarchy (DR2).

The overview dendrogram (Figure III.3(a)) shows all clusters in the hierarchy, from the root down to a user-chosen level. We do not display nodes below this level to reduce the visual complexity of the exploration process. The leaves in the resulting dendrogram visualization show the spatial situations that are representative for the respective clusters. These spatial situations are depicted as cartographic maps that encode the data's scalar values with diverging or sequential color schemes, depending on the data. To obtain a cluster representative, we compute an average spatial situation

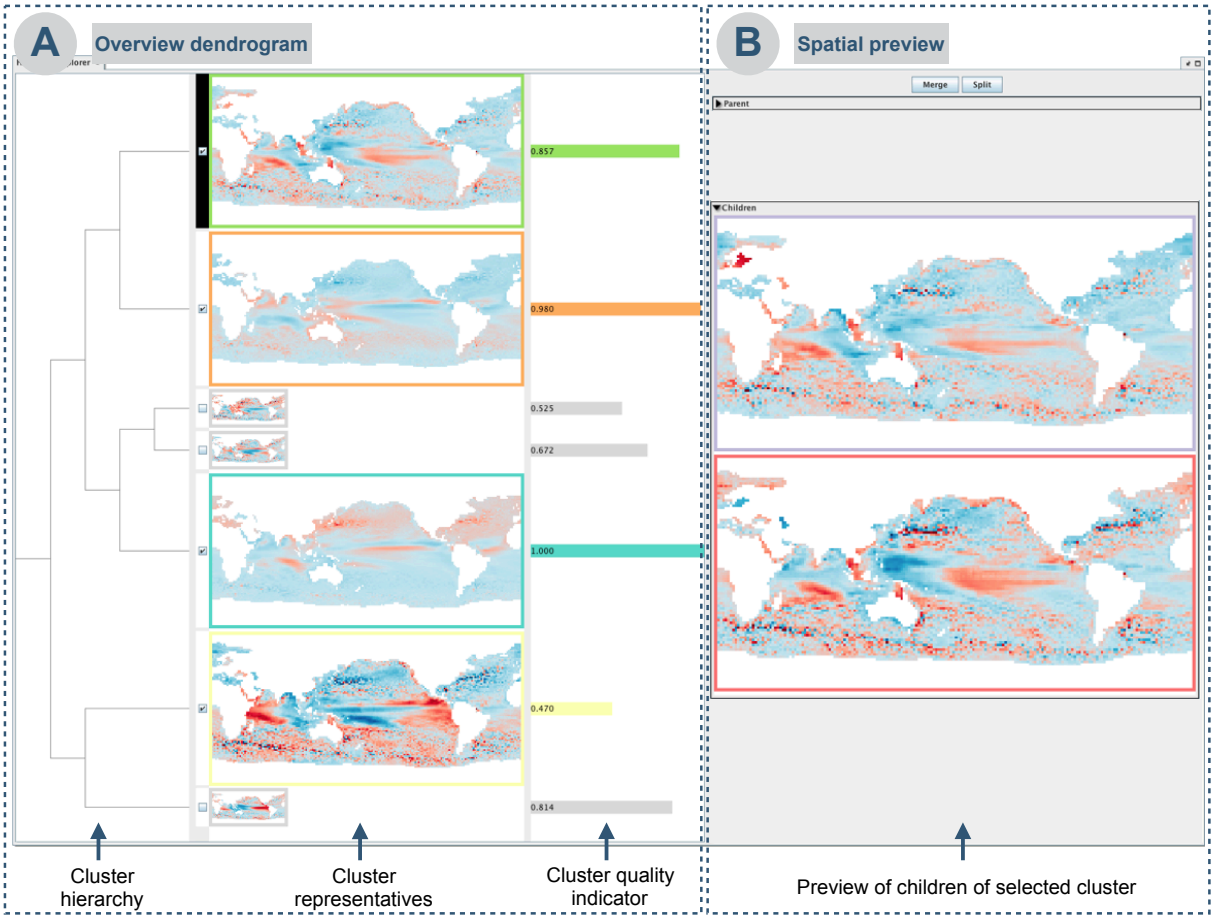



**Figure III.3:** Hierarchy explorer and its two main components: (a) overview dendrogram and (b) spatial preview.

from all spatial situations in a cluster. The representative *Rep* of cluster $C$ is an $N \times M$ matrix with $Rep = \frac{1}{l} \sum_{i \in C} D_i$ with $l = |C|$ and $D_i$ is the spatial situation associated to time step $i$. To provide hints on how well a cluster representative portrays the entire group of associated time steps, each cluster representative has an associated horizontal bar whose length encodes the average dissimilarity of all time steps in a cluster to its representative spatial situation. Cluster representatives with small average dissimilarities are usually a good approximation of the underlying time steps. The vertical alignment of leaves and their associated horizontal bars in the proposed visual encoding facilitate intercluster comparison. Scientists may use this information to focus on cluster representatives, further refining them where necessary. To remove extracted clusters temporarily from the visual summary, users can deselect leaves in the overview dendrogram. This reduces their size to thumbnail images. Additionally, deselected clusters do not appear in the spatial gallery and are grayed out in the sequence view and the periodicity explorer. Domain experts can further manually rearrange the hierarchy if they consider two clusters that are located in distant parts of the dendrogram as similar. They can

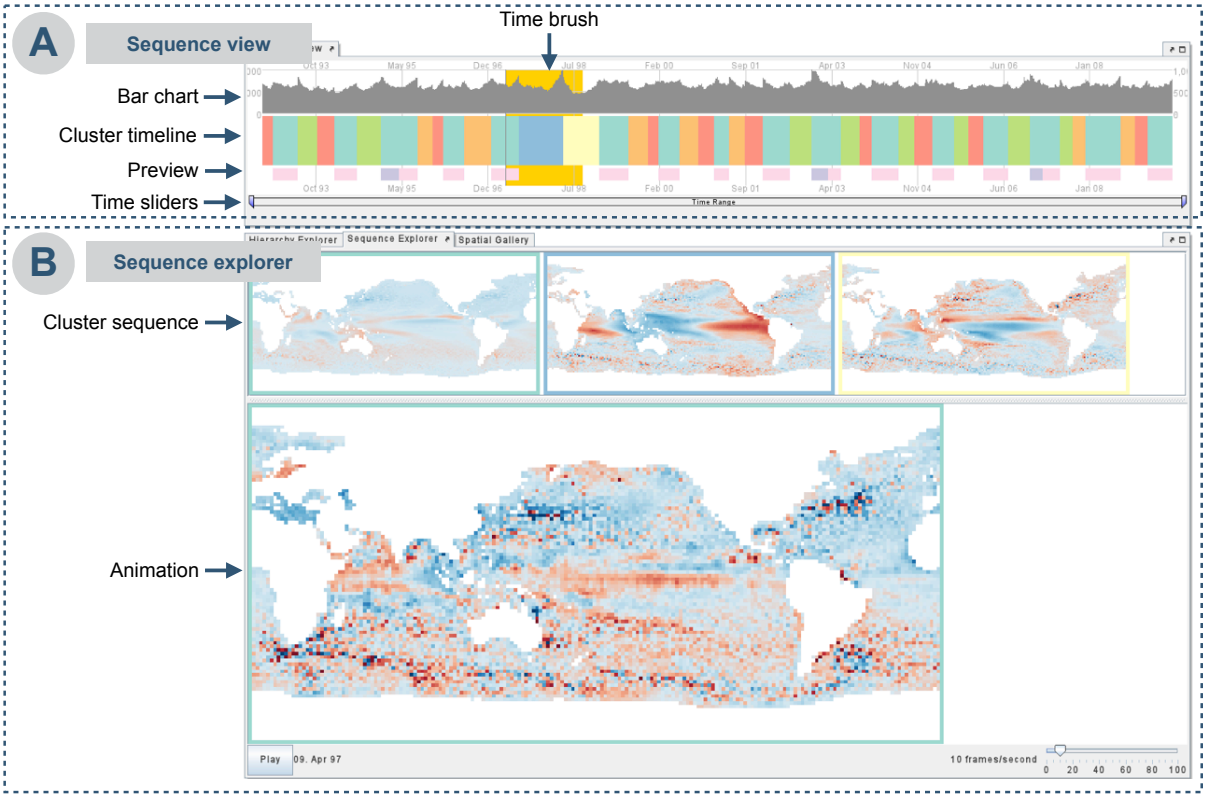



**Figure III.4:** (a) Sequence view and (b) sequence explorer.

easily merge two such clusters via drag-and-drop. The subsequent reba- lancing of the dendrogram is illustrated in Figure III.1.

For any cluster selected in the overview dendrogram, the spatial preview (Figure III.3(b)) allows users to see the representatives of its two child clusters. After assessing the spatial patterns, users can drill down the dendrogram to split the selected cluster. Alternatively, they can either roll up the dendrogram and merge the selected cluster with its sibling or focus on a different cluster. The overview dendrogram and the spatial preview are vertically scrollable if screen space does not suffice.

### Sequence view

The sequence view comprises the following three parts (Figure III.4(a)): a cluster timeline that presents the temporal order of clusters in a visual summary (DR1), a preview that supports gradual refinement of a user-selected cluster by showing potential changes in the temporal information (DR2), and a bar chart that provides hints on the quality of a visual summary (DR4).

The cluster timeline is a color-coded horizontal bar that represents the entire time series. Cluster affiliation of time steps is mapped to color. The bar chart above the cluster timeline depicts for

each time step its dissimilarity to the associated representative spatial situation. To compute the dissimilarity, we use the same measure that was applied in the hierarchical clustering. Small values in the bar chart indicate time steps where the visual summary fits well; high values point to parts where a cluster representative is not an adequate description of individual time steps. Upon interactive selection of a cluster in the cluster timeline, the preview appears below the timeline. It presents a smaller masked version of the cluster timeline that contains only time steps that belong to the selected cluster and, thus, also visually marks the currently selected cluster. The time steps in the preview are color-coded as if the selected cluster was split into its two children. In addition, horizontal sliders allow users to zoom in time. A time brush lets users select time periods of interest for further inspection in the sequence explorer.

**Sequence explorer**

The sequence explorer (Figure III.4(b)) contains two components that, for a user-selected time period, allow scientists to assess the spatiotemporal dynamics in the data (DR1) and compare the original data with the cluster representatives (DR4). In the horizontally scrollable cluster sequence, the representative spatial situations of clusters are arranged from left to right as they occur in the sequence view. Colored frames around the cluster representatives encode cluster affiliation. The animation depicts the original data. For small time periods, looking at the original data in an animation allows users to qualitatively evaluate the spatiotemporal dynamics in the data and assess the representativeness of certain patterns in the visual summary. We provide several visual aids to help users analyze the animated sequence. The cluster affiliation of the time step that is currently depicted in the animation is encoded in a colored frame around the animation window. In addition, the animation is visually linked to the sequence view. A vertical line in the sequence view locates the current time step, and a colored rectangle represents the selected time period.

**Periodicity explorer**

The periodicity explorer (Figure III.5) helps scientists to visually analyze the data for various types of recurring behavior (DR5). It splits the cluster timeline into intervals of equal length. These intervals are then chronologically arranged in rows from top to bottom, resulting in a 2D array. Users can freely determine the interval length and interval start date. Since cluster affiliation is mapped to color, recurring phenomena become visible when the same color appears in multiple rows in roughly the same horizontal position.

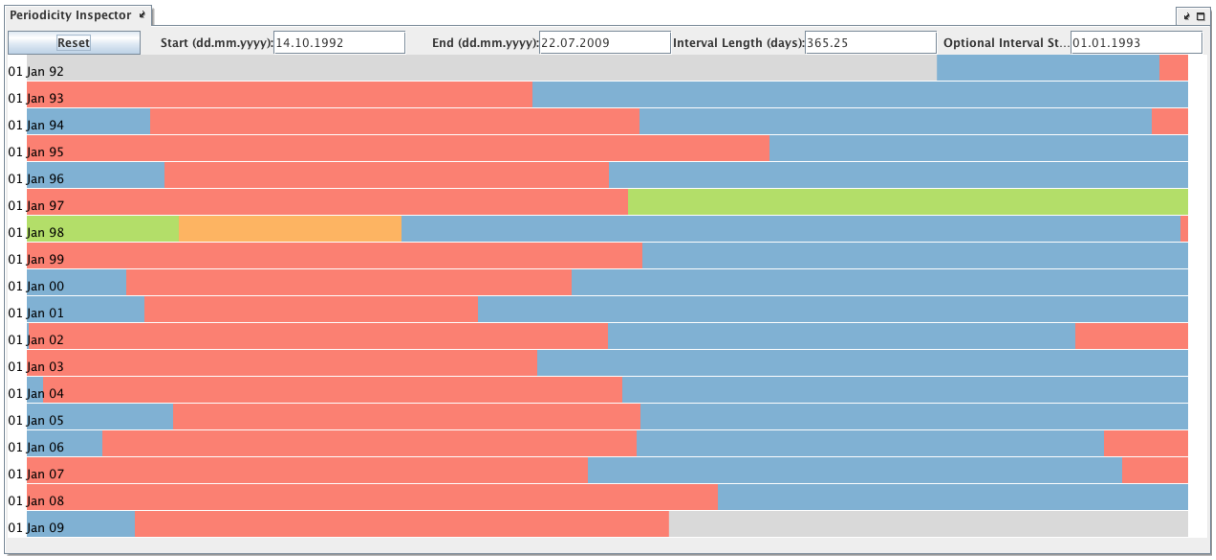



**Figure III.5:** Periodicity explorer arranges the cluster timeline in a two-dimensional array. Users can choose any interval length and interval start date to visually detect (quasi-)periodic behavior. This particular example shows the periodic Winter-Spring (red)/Summer-Fall (blue) cycle that can be observed in global sea surface height observations from 1992 to 2009. The outstanding green and orange clusters represent a very strong El Ni  no event in 1997/1998.

### Spatial gallery

To present the (intermediate) results of the exploration process and allow users to compare the extracted spatial patterns (DR1), the spatial gallery provides a maximum of screen space to depict the cluster representatives (Figure III.6). It arranges the spatial situations in a matrix layout. Users can adjust the size of the spatial patterns by choosing the number of columns in the matrix. The gallery is vertically scrollable if screen space does not suffice.

### Color coding

In our tool, the different views are visually linked by a consistent color coding of clusters. We assign a unique color to each cluster that users visit in the hierarchy. This approach requires a high number of distinguishable colors. To this end, we use one of ColorBrewer's qualitative color schemes [52] as well as colors sampled from the CIELAB color space (please see Guo et al. [35] for a suitable sampling strategy). We chose the ColorBrewer colors to provide users with a carefully designed and easily distinguishable color scheme. If the exploration process requires additional colors, we use the CIELAB samples. This strategy yields a sufficient number of distinguishable colors. In addition, users can change the colors manually to adjust the color coding according to their preference.

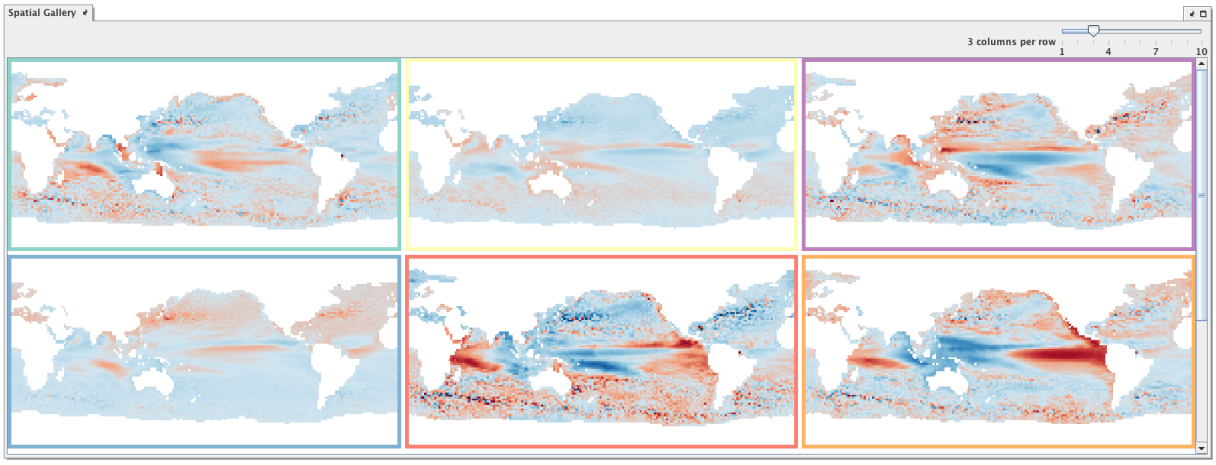



**Figure III.6:** Spatial gallery depicts the cluster representatives of the current visual summary in a matrix layout. The slider allows users to choose the number of columns in the matrix.

### Scalability

Two main factors influence the scalability of our tool: the number of time steps in the geospatial time series and the number of clusters extracted from the cluster hierarchy.

Our tool can display a large number of time steps, as long as the cluster affiliation of subsequent time steps yields visually coherent blocks in the sequence view. Regarding the number of clusters extracted from the cluster hierarchy, we observed that geoscientists focus on a rather small subset of patterns when exploring geospatial time series. They normally analyze between 10 and 30 clusters. Therefore, we specifically designed the hierarchy explorer to support such a focused exploration. Technically, the hierarchy explorer can depict a larger number of clusters due to its vertically scrollable components.

## 5.3 Application example

One cornerstone in the development of our tool was the intense collaboration with ocean modelers. The goal was to help ocean modelers in the analysis of observational data as well as output from model simulations.

We have already presented the initial results of our collaboration in our previous study [81] where we demonstrated that a static visual summary can capture various characteristic spatiotemporal patterns in geospatial time series. A static version of a visual summary, however, does not allow scientists to gain a more detailed understanding of the presented patterns. Scientists need to be able to extract different sets of spatial situations from the data and assess the corresponding visual summaries

interactively to focus on patterns of interest and further differentiate these patterns into subtypes. Concurrently, they need to be able to eliminate other patterns that they consider insignificant or distracting.

Here, we present an example of how scientists used our interactive tool to identify and further distinguish different types of El Niño events. We also give a short example of how our tool helped scientists generate hypotheses about processes in the ocean.

## 5.4 Identification and differentiation of El Niño events

To evaluate our tool, one of our collaborators used it to identify and differentiate known El Niño events in ocean observational data. The data used were daily satellite observations of sea surface temperatures (SSTs) in the Tropical Pacific, including smaller sections of the Caribbean and Southeast Asia, covering a time period from 2000 to 2010. Since the seasonal cycle dominates the variability of the time series without being of interest for the present study, we deseasonalized the data by subtracting climatological monthly means. To focus on interannual variability, we further computed weekly mean SSTs. The result of these pre- processing steps was a time series of SST anomalies with 574 time steps and a spatial resolution of $661 \times 240$ grid points.

The prominent processes in the described region are El Niño and La Niña events. El Niño events are irregularly (about every 2 to 7 years) recurring phases of anomalously high SSTs in the Tropical Pacific with implications on weather patterns worldwide. La Niña events, on the other hand, are characterized by cold SSTs in the Tropical Pacific. Two different El Niño types can be observed: An Eastern Pacific El Niño shows maximum positive SST anomalies close to the Ecuadorian and Peruvian coast; a Central Pacific El Niño shows the strongest positive SST deviations close to the date line. Occasionally, there are positive anomalies in both places; some authors have therefore defined a third, "mixed" type [154].

Since ocean scientists have identified three Central Pacific El Niños and one Eastern Pacific El Niños in the observed region between 2000 and 2010,46 the task was to correctly locate and distinguish these events in the SST data. Our collaborator used our tool to create a coarse visual summary of the time series that describes the data with only two clusters. These two clusters already revealed the two dominant processes in the region (Figure III.7). One cluster represented time steps that were somewhat influenced by La Niña, while the other represented time steps that were more influenced by El Niño. Further exploration focused on the latter. Relatively high SST values along the Equator in a cluster representative hint at El Niño events in this cluster. After selecting such a

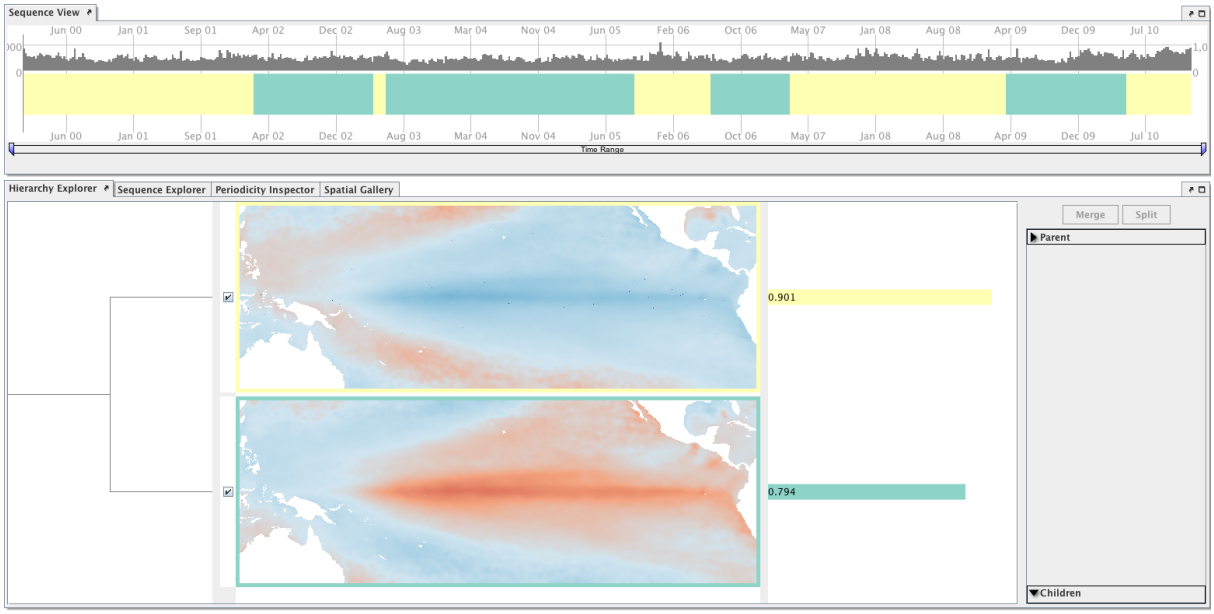



**Figure III.7:** Initial stage of the exploration of sea surface temperature anomalies. Splitting the data into two clusters reveals the two dominant processes in the region. The yellow cluster represents time steps that are somewhat influenced by La Niña, while the green cluster represents time steps that are more influenced by El Niño.

cluster and examining the spatiotemporal patterns of its two children, the ocean modeler decided whether splitting the selected cluster would reveal relevant information. Decisions to split or merge particular clusters were grounded on the information that our tool provides about the quality of the visual summary during the exploration process and on his knowledge about what typical El Niño situations look like. The goal was to separate El Niño phases from adjacent, rather neutral, phases that were assigned to the same cluster. After several split operations, our collaborator was able to identify all four El Niño events: three Central Pacific types and the only Eastern Pacific El Niño (Figure III.8). To gain additional confidence in the result, he selected time periods associated with these events in the sequence view and examined the original data in the sequence explorer.

## 5.5 Discussion

In the described application example, the scientist was able to identify all three known Central Pacific El Niño events and the only Eastern Pacific El Niño. Although one would expect all three Central Pacific El Niño events to be in the same cluster, Figure III.8 shows that the 2004/2005 event is located in a separate cluster. It is represented by the purple cluster, while the other two Central Pacific El Niño events are represented by the blue cluster. Our collaborator explained this with the event's extremely low intensity. Please note that our tool was not specifically tailored to the identification

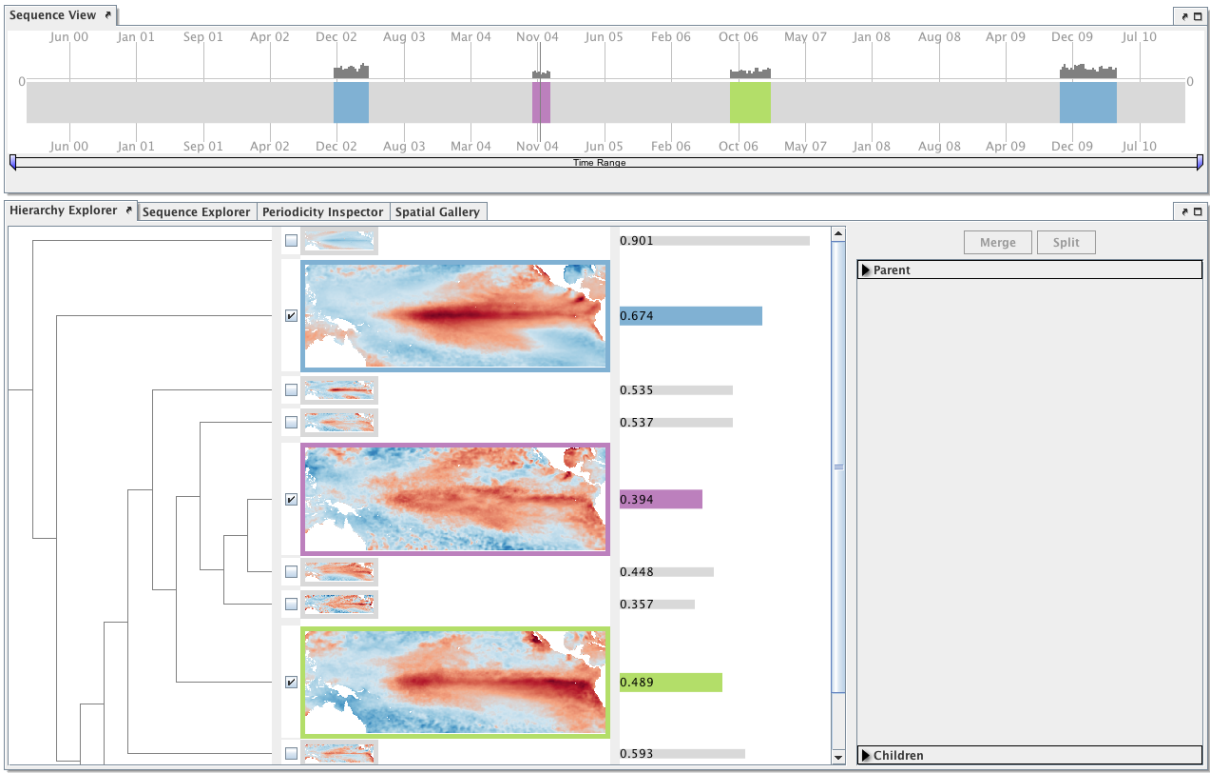



**Figure III.8:** Result of the exploration of sea surface temperature anomalies. After several split operations, our collaborator was able to identify all four El Niños. Two out of the three Central Pacific El Niños are represented by the blue cluster, and the other one is represented by the purple cluster. The only Eastern Pacific El Niño is depicted by the green cluster.

and differentiation of El Niño events. Therefore, it is encouraging that such detailed patterns could be distinguished with our general purpose approach. Geoscientists normally combine a variety of methods to detect these events, for example, regression analysis, empirical orthogonal functions, and wavelet analysis [69]. They also often focus in their analysis on various indices that describe particular geographic regions with respect to a specific environmental process [154]. In contrast, our approach makes very few assumptions about the data and allows scientists to analyze a variety of patterns for large geographic regions without having to refer to specialized indices. Once our tool has pointed experts to interesting patterns, they can apply established quantitative methods for further testing and inspection.

Overall, our collaborators valued the intuitiveness of the interactive exploration. They appreciate the ability to progressively increase the level of detail of spatiotemporal patterns in the hierarchy explorer and value the permanent link to the corresponding original data in the sequence explorer. They confirmed that the tool supports detection of characteristic patterns as well as differentiation into their subtypes.

After applying our tool to different data in their daily research, scientists pointed out that it allows them to produce hypotheses. In one particular example regarding satellite observations of sea surface heights, our tool suggested a seasonal cycle in a geographic region where experts did not expect it. Our tool pointed scientists to this particular feature in the data, which will now be a starting point for further investigation.

Our collaborators also shared their thoughts on potential limitations. Regarding the hierarchical clustering, it became apparent that sometimes a characteristic spatial situation is represented by several clusters on different branches in the dendrogram. This leads to redundant spatial situations in the visual summary. Here, our collaborators appreciated the ability to manually rearrange the hierarchy. The analysis of geospatial time series with very low temporal autocorrelation can be quite challenging with our tool. The resulting visual summaries are difficult to interpret since the cluster affiliation of subsequent time steps changes frequently and, thus, the sequence view does not display visually coherent blocks. To address this problem, scientists may use the periodicity explorer to create visually coherent blocks by rearranging the cluster timeline in a two-dimensional array. This distributes the cluster timeline across multiple rows, providing more screen space. In addition, visually coherent blocks may not only become apparent along the horizontal axis, but also become apparent along the vertical axis. Another option is a more substantial preprocessing of the data, for example, by temporal or spatial filtering, to remove processes that are not of interest and that can be theoretically estimated.

## 6 Conclusion and future work

Close collaboration with geoscientists enabled us to identify and address a major challenge in geospatial time series analysis: the complexity of the processes described in the data, which requires a simultaneous assessment of the data's spatial and temporal variability. To address this challenge, our approach supports users in the analysis of geospatial time series by extracting different sets of spatial situations from the data and exploring the corresponding visual summaries. We use the output of agglomerative hierarchical clustering of time steps as a starting point for interactive visual exploration. A thorough task analysis allowed us to elicit appropriate design requirements for the visual exploration tool. The tool comprises five visualization components that each focus on different aspects of the interactive analysis.

We received detailed user feedback at every stage of the development process, refining our ap-

proach with every iteration. Our tool is currently applied by geoscientists in their daily research. Their feedback suggests that it allows them to explore the data for a great variety of processes and patterns, leading to new hypotheses and eventually generating new scientific insight. The next challenge is to evaluate our approach in a longitudinal user study to gain an understanding of the conceptual limitations of our approach and identify roads for improvements in the visual encoding and analytical interaction.

We have identified several research directions to extend our approach. Since the segmentation of the data by means of clustering can be regarded as a symbolic representation of the time series, we plan to include motif mining techniques to facilitate automatic detection of periodicities and compute even more compact visual summaries of geospatial time series. Furthermore, we want to support comparison of clustering results for different geospatial time series. Finally, we would like to extend our approach to multirun simulation output. This is a challenging task since the visual encoding of multirun data is an open research question. We hope that building a visual exploration tool for multirun data will contribute to a better understanding of simulated processes in many geoscientific application scenarios.

## Acknowledgements

The authors thank Tobias Rawald and Ralf Friedeman for their help in implementing the prototype. This study was partially supported by the German Federal Ministry for Education and Research (BMBF) via the Potsdam Research Cluster for Georisk Analysis, Environmental Change and Sustainability (PROGRESS) (grant number 03IS2191A).

# Chapter IV

# Visual analytics for comparison of ocean model output with reference data: detecting and analyzing geophysical processes using clustering ensembles

**Published as:** P. Köthur, M. Sips, H. Dobslaw, and D. Dransch. Visual Analytics for Comparison of Ocean Model Output with Reference Data: Detecting and Analyzing Geophysical Processes Using Clustering Ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1893–1902, doi:10.1109/TVCG.2014.2346751, December 2014.

**Abstract.** Researchers assess the quality of an ocean model by comparing its output to that of a previous model version or to observations. One objective of the comparison is to detect and to analyze differences and similarities between both data sets regarding geophysical processes, such as particular ocean currents. This task involves the analysis of thousands or hundreds of thousands of geographically referenced temporal profiles in the data. To cope with the amount of data, modelers combine aggregation of temporal profiles to single statistical values with visual comparison. Although this strategy is based on experience and a well-grounded body of expert knowledge, our discussions with domain experts have shown that it has two limitations: (1) using a single statistical measure results in a rather limited scope of the comparison and in significant loss of information, and (2) the decisions modelers have to make in the process may lead to important aspects being overlooked.

In this article, we propose a visual analytics approach that broadens the scope of the analysis, reduces subjectivity, and facilitates comparison of the two data sets. It comprises three steps: First, it allows modelers to consider many aspects of the temporal behavior of geophysical processes by

conducting multiple clusterings of the temporal profiles in each data set. Modelers can choose different features describing the temporal behavior of relevant processes, clustering algorithms, and parameterizations. Second, our approach consolidates the clusterings of one data set into a single clustering via a clustering ensembles approach. The consolidated clustering presents an overview of the geospatial distribution of temporal behavior in a data set. Third, a visual interface allows modelers to compare the two consolidated clusterings. It enables them to detect clusters of temporal profiles that represent geophysical processes and to analyze differences and similarities between two data sets.

This work is the result of a close collaboration with ocean modelers. They employed our concept to find aspects of improvement in a new version of the Ocean Model for Circulation and Tides.

# 1 Introduction

Geoscientific simulation models, in particular ocean and climate system models, consider complex interactions between processes in the Earth system [137]. For example, the salinity of the oceans affects the circulation within oceans, which in turn impacts the energy exchange between oceans and atmosphere. The latter influences air temperature, which may affect continental and sea ice. To come full circle, melting ice leads to freshwater influx into the oceans which influences their salinity.

Modeling such interactions serves three important purposes. First, simulation models provide data about real-world phenomena in geographic regions that are not or only partially covered by monitoring devices. Second, they enable scientists to identify and study causalities between geophysical processes to gain insight into the physics of the Earth system. Last, when researchers have acquired a sound understanding of the mechanisms within the Earth system and their interactions, they can produce reliable predictions; a prominent example being future greenhouse gas concentrations and their effect on, e.g., sea level rise, availability of fresh water, or natural hazards.

In this paper, we specifically focus on ocean models. Intense collaboration with ocean modelers at the German Research Center for Geosciences GFZ enabled us to gain an understanding of the challenges involved in model assessment. The development of ocean models is an iterative process in which the assessment of new model versions is a critical part. After each change to the model, researchers compare the model output to reference data to determine whether the new version improves the simulation. The reference data may either stem from a previous version of the same model or from observations. The latter is usually the case when a model is new and there is little knowledge about its behavior and performance against observations. The former applies, e.g., when scientists want to improve specific aspects of a tried-and-tested model and are already familiar with the observational data and the behavior of previous model versions.

For the comparison, modelers need to locate and compare geophysical processes in both data sets. In our context, a geophysical process is a broad concept. It includes, e.g., the El Niño southern oscillation or particular ocean currents such as western or eastern boundary currents. Every process has particular temporal and geospatial characteristics that manifest to a varying degree in the data.

To detect geophysical processes, our partners try to identify and locate temporal profiles in the data that are characteristic for the processes under study. This is a challenging task due to the volume and complexity of the data. Ocean models depict the ocean as a regularly structured three-dimensional grid. The grid points represent geographic coordinates and have temporal profiles

associated with them that describe the temporal behavior at these coordinates. For the analysis, modelers frequently focus on the topmost layer, the sea surface. This is appropriate because most mechanisms within the ocean manifest themselves in changes to sea surface heights. But even with the focus on the sea surface, the data comprise thousands or hundreds of thousands of time series.

To address this challenge, scientists employ data aggregation and visualization in a two-step process. First, modelers aggregate the time dimension by computing a single statistical measure for each temporal profile in the data. Scientists plot this measure in a separate geographic map for each data set to visually analyze and compare its geospatial distribution. Second, modelers choose a small number of geographic coordinates (typically not more than 20) for a more detailed analysis of the temporal behavior. Line charts are used to study and compare the temporal profiles associated with the selected coordinates.

Although scientists aggregate and compare the data based on experience and a well-grounded body of expert knowledge, they know that data aggregation results in loss of information, and that the subjectivity involved in this strategy may have them miss important aspects. The chosen statistical measure focusses the analysis on a particular characteristic of temporal behavior; information about other aspects of temporal behavior is lost. In addition, other important details may not be noticed because modelers focus the comparison of temporal profiles on a limited number of geographic coordinates.

In this article, we introduce a visual analytics approach that presents modelers with a more comprehensive view on the temporal behavior in the data. It allows modelers (a) to create multiple spatial clusterings of the temporal profiles in model output and reference data, (b) to consolidate the various clusterings for each data set with an ensemble approach [131], and (c) to interactively explore and compare the two consolidation results.

We chose clustering of temporal profiles because (1) it allows modelers to systematically identify and locate the predominant types of temporal behavior in the data, and (2) because it allows researchers to base the analysis on many different characteristics of temporal behavior. Our concept enables scientists to compute multiple clusterings with varying user-chosen features of temporal behavior, clustering algorithms, and parameterizations. In the following, a *feature* denotes any representation of a temporal profile that captures a particular aspect of temporal behavior and supports definition of a (dis)similarity metric. To unite the different aspects of temporal behavior reflected in various clusterings, a clustering ensemble combines all clusterings of one data set into a single consolidated clustering. A visual interface facilitates comparison of the two consolidation results for

model data and reference data. It allows researchers to identify clusters of temporal profiles that represent geophysical processes as well as to explore differences and similarities between the two data sets.

In particular, the contributions of this article are as follows:

- We closely collaborated with ocean modelers and conducted a thorough task and requirement analysis to identify the key challenges in the comparison of ocean model output with reference data.
- We combine cluster ensembles and interactive visual exploration for a novel approach to supporting the assessment of ocean models.
- We demonstrate how our concept enables modelers to conduct a fast comparison of model data with reference data that complements the existing statistical methods.

## 2 Related work

Although numerous guidelines and techniques exist for the visual analysis of geospatial data [39,94], time series data [2], and geospatio-temporal data [9,11], the complexity and volume of geoscientific data still presents significant challenges [90]. Clustering is an established technique for approaching these challenges. Its aim is to divide data into groups of similar objects (for further reading, please refer to [15, 50]). A number of visual analytics works apply clustering to analyze geospatial time series. Andrienko et al. [7] introduce two perspectives to such an analysis: 'space-in-time' and 'time-in-space'. The former analyzes how the geospatial distribution of data values changes over time; the latter studies how the temporal behavior is distributed in geographic space. To address both perspectives, the before mentioned work uses self-organizing maps (SOM) as a clustering and visualization technique and combines it with multiple linked views. In own previous work [82], we consider the 'space-in-time' perspective and combine hierarchical clustering with visual exploration to support detection of dominant spatial states in geoscientific data. Approaches that apply clustering to analyze multiple temporal profiles ( 'time-in-space') are more numerous. Guo et al. [35] and Andrienko et al. [6] use SOMs and combine them with geographic maps, small multiples, reorderable matrices, time series charts, or parallel coordinates. Another work by Andrienko et al. [10] combines clustering and interactive visual analysis with the aim of statistical modeling of geospatial time series. Woodring and Shen [151] combine wavelet transform, clustering, and interactive visualization

for analysis of trends at varying temporal scales. These works use a single clustering, which captures only a particular aspect in the data. A different clustering leads to different results. In our application, however, it is important to consider multiple aspects of temporal behavior simultaneously.

Clustering ensembles [159] address this issue. They combine multiple clusterings into one clustering solution that shares as much information as possible with the input clusterings. With this approach one is able to cluster the data with varying features. In addition, it eases the burden on the users to find an optimal combination of (dis)similarity measure, algorithm, and parameterization of the clustering. They can select a set of plausible configurations and use cluster ensembles to combine the results. The resulting consolidated clustering is generally more robust and more accurate [47, 107, 131]. Although clustering ensembles have been shown to improve data analysis in a variety of fields – e.g., cancer research [66, 155] or remote sensing [157] – we are, to the best of our knowledge, not aware of any works that apply this approach and its benefits to ocean modeling.

Another important aspect of our concept is the visual comparison of model data with reference data. As noted in a recent survey [90], many approaches for visualization and visual analysis in the Earth sciences have been introduced, but only a few works support the comparison of geoscientific simulation data. Nocke et al. [108] provide a library of comparative visualization techniques tailored to climate modeling. Based on the characteristics of the data and the task at hand, their framework generates an appropriate visualization. Unger et al. [142] address the validation of geoscientific simulation models. They compare many model outputs with sparse and uncertain observations. The focus is to find an appropriate model parameterization that best matches the observations. Ahrens et al. [1] use comparative visualization to support detection of errors in simulation model code. To this end, they conduct a numerical comparison of several output variables. Kehrer et al. [74, 75] and Ladtstädter et al. [85] support visual analysis and comparison of different variables of climate model output by multiple linked views. Recent work by Poco et al. [113] focusses on the comparison of output from different climate models. Their approach concentrates on the analysis of correlations between data sets. These works do not base the comparison on geophysical processes, a key requirement of our users.

The concept that is closely related to our application, focusing on processes and applying clustering, was introduced by Frey et al. [48]. They support comparison of two temporal field data sets by combining automated detection of processes with interactive visual exploration. For the detection of processes, they use recurrence analysis and clustering. This approach regards processes as recurring events and places the emphasis on temporal similarity between data. In our application, we do not

focus on recurrences but are interested in geographic regions that exhibit similar temporal behavior.

# 3 Visual analytics approach and requirements

We adopted a user- and task-centered approach [37] in our collaboration with ocean modelers at the German Research Center for Geosciences GFZ. This involved frequent meetings and discussions with our partners to obtain a detailed understanding of the model assessment process and the associated challenges. In this section, we provide an overview of our concept and the associated requirements.

## 3.1 Objectives for a visual analytics approach

As a result of our analysis we identified three main objectives for a visual analytics approach to facilitate comparison of model data with reference data.

**(1) Less temporal aggregation** Using a single statistical measure as a feature to describe a time series is a rather drastic approach to temporal aggregation. The ability to employ other types of features, such as the power spectrum of a time series, would reduce the amount of information lost and allow modelers to study more sophisticated characteristics of temporal behavior.

**(2) More comprehensive comparison process** The current comparison process requires modelers to make two main decisions that are rather subjective and may result in important aspects to remain hidden in the data. First, they have to choose a feature for temporal aggregation. A single feature, however, only describes a particular aspect of the temporal behavior. Other features may capture different – but equally valid – aspects, and, hence, may yield different results. Being able to consider various features of temporal behavior in the comparison would broaden the scope of the analysis. Second, modelers hand-pick geographic coordinates for detailed comparison of temporal behavior. However, there is no guarantee that all relevant behavior can be observed at the selected coordinates. To reduce the risk of overlooking important aspects, modelers need to take geographic areas into account, not just a few coordinates.

**(3) Enhanced visual exploration and comparison** When the two objectives above are met, modelers will be able to study the output of a model in comparison to reference data from a broader perspective. To take full advantage of the additional information, scientists need a visual analytics

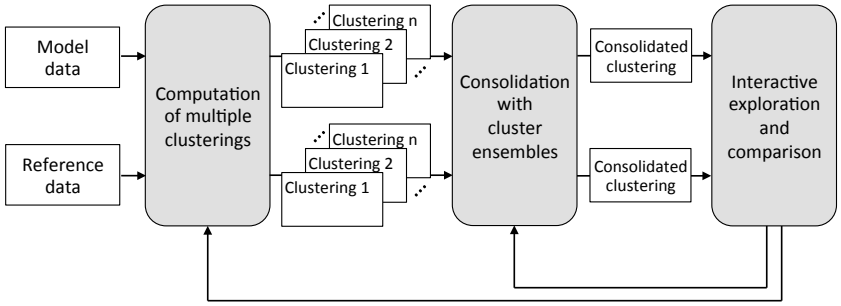



**Figure IV.1:** Our visual analytics concept: (1) modelers cluster model data and reference data with multiple configurations, (2) each set of clusterings is combined into one consolidated clustering, and (3) modelers interactively explore and compare the two consolidated clusterings. At any time during the exploration, scientists can either change the configuration of the consolidation or the set of input clusterings. Note that the two data sets are considered independently in the clustering and consolidation process.

tool that supports the comparison from this broader perspective. The tool should enable modelers to quickly identify and inspect temporal profiles that point to relevant geophysical processes in model data and reference data, and to assess differences and similarities between the data sets.

## 3.2 Concept

Based on our analysis and the identified objectives, we developed a threefold concept (Figure IV.1):

1. *Computation of multiple geospatial clusterings.*

   Our concept allows modelers to widen the scope of the comparison by performing multiple clusterings of the temporal profiles in model data and reference data. Since clustering considers the entire geographic domain and systematically detects groups of similar temporal profiles; modelers can compare geographic regions (the clusters) instead of a few geographic coordinates. For the clusterings, scientists can choose from a broad range of features of temporal profiles, instead of just single statistical measures. This provides them with more options for the detection of geophysical processes. The resulting two sets of clusterings represent various perspectives on the temporal behavior in the data.

2. *Consolidation of the clusterings via clustering ensembles.*

   To enable modelers to compare model data and reference data based on the two sets of clusterings, we combine each set into a separate consolidated clustering using clustering ensembles [131]. From a modeler's perspective, the consolidation of clusterings that are based on different features presents a more comprehensive, more robust view on the temporal behavior

in a data set.

3. *Interactive visual exploration and comparison of the consolidated clusterings.*

   An interactive visual interface allows modelers to study and compare the two consolidated clusterings of model data and reference data. Scientists can perform visual queries to identify clusters that represent geoscientific processes, and to explore differences and similarities between the two data sets with respect to these clusters. Furthermore, modelers can go back to the previous two steps in the pipeline at any time. They can either choose to explore a differently configured consolidation result or they can change the set of input clusterings to, e.g., consider additional features of temporal behavior.

## 3.3 Requirements

In discussions with ocean modelers, we were able to identify four analytical requirements (ARs) and three visualization requirements (VRs) for our concept.

### Analytical requirements

The analytical requirements comprise two requirements for the computation of multiple clusterings (AR1 and AR2), and two requirements for the consolidation with clustering ensembles (AR3 and AR4).

AR1 *Broad range of features*

In order to detect different geophysical processes or to study various characteristics of a single process, the set of input clusterings must represent different kinds of temporal behavior. Therefore, modelers need to be able to choose from many features to capture a wide variety of characteristics of temporal behavior when computing the geospatial clusterings.

AR2 *Many distinct cluster parameterizations*

A parameterization of a clustering algorithm reflects a specific assumption about the data to be clustered. However, modelers can often only make vague assumptions about the temporal behavior in model and reference data. For example, although the number of distinct types of temporal behavior in a data set can usually be narrowed down to a plausible range, it is difficult to anticipate the exact number. To account for this challenge, modelers need to be able to conduct clusterings with a varying number of clusters.

AR3 *Flexible configuration of the consolidation*

Modelers must consider two important aspects of a consolidated clustering: the amount of information it shares with the set of input clusterings, the quality, and its number of clusters, the complexity. A consolidated clustering that shares little information with the set of input clusterings is difficult to interpret. Likewise, a large number of clusters also complicates interpretation and comparison. To achieve a good balance between quality and complexity, modelers need to be able to configure different consolidations with a varying number of clusters and to study and compare the results.

AR4 *Quantitative measures to support the assessment of consolidated clusterings*

When presented with a consolidated clustering, modelers want to know how much information the individual clusters share with the input clusterings. This enables them to discriminate the clusters whose geographic locations were often considered as similar in the input clusterings, from the clusters where the input clusterings agree less on. This is important since the latter may not allow for a meaningful interpretation. Furthermore, to support an initial assessment of the relationships among clusters, modelers need quantitative measures that describe the geospatial similarity and the feature similarity between consolidation clusters.

**Visualization requirements**

VR1 *Overview of consolidated clusterings*

Modelers have to be able to interpret the results of the cluster ensembles. Therefore, they need to obtain an overview of the clusters and their relations in the two consolidated clusterings for model data and reference data. This requires visualizations that allow them (a) to filter the clusters that share only very little information with the input clusterings, (b) to detect clusters in each data set that may represent relevant geophysical processes, and (c) to identify potential matches between clusters from different data sets to guide further comparison. This requirement is associated with the following tasks:

- Assess distribution of clusters in geographic space.
- Obtain overview of geospatial similarity among clusters.
- Obtain overview of feature similarity among clusters.
- Compare individual clusters regarding the information they share with the input clusterings.

VR2 *Inspection of cluster properties*

Modelers must understand the properties of a single consolidated cluster to judge whether it is related to a geophysical process. To this end, they need to:

- Assess the temporal variations of input time series associated with a cluster.
- Inspect the distributions of feature values associated with a cluster.
- Inspect the distribution of cluster members in geographic space.
- Explore relations between input time series, feature values, and geographic distribution.

VR3 *Detailed comparison of clusters*

For a judgement of model quality, modelers need to explore and evaluate differences and similarities between consolidated clusters in model data and reference data. This allows them to identify geographic regions where the model performs well and where it needs improvement. The associated tasks are:

- Compare the temporal variations of input time series associated with clusters.
- Compare the distributions of feature values for multiple clusters.
- Compare the distribution of clusters in geographic space.

# 4 Clustering and consolidation

In this section, we describe the clustering and consolidation part of our concept and how it meets the analytical requirements AR1–AR4.

## 4.1 Computation of multiple geospatial clusterings

The computation of multiple geospatial clusterings has two aspects. As explained in AR1 (Section 3.3), modelers need to be able to cluster the data based on a variety of features. Therefore, our approach provides many features describing aspects of temporal behavior that scientists consider important for detecting geophysical processes; e.g., minimum and maximum value, mean, standard deviation, power spectrum, and logarithmic power spectrum. Note that scientists have the choice to cluster the raw data without prior computation of features. Scientists can also add additional features to focus on other types of geophysical processes.

Secondly, modelers can create multiple clusterings by varying the parameterizations of a clustering algorithm (AR2). The challenge in our application scenario was to identify a clustering method that is appropriate for ocean model output and reference data. We conducted a large number of experiments, clustering well understood observational data with different algorithms, parameterizations, features, and distance or similarity measures. The methods tested in these experiments were hierarchical clustering [71], DBSCAN [43] as a density-based method, a Gaussian mixture model approach [46], and k-means [12]. The clusterings were conducted with the following distance and similarity measures: Euclidean, Manhattan, mutual information [29], normalized compression distance [14], dissimilarity based on cross-correlation [88], and dynamic time warping [16]. We chose k-means and Euclidean distance because this combination yielded meaningful clusters over a broad range of parameterizations and features. Note that our concept is not limited to a specific clustering algorithm or distance measure. If need be, other methods can be included to provide modelers with additional options.

To satisfy AR2, modelers can vary the number of clusters and the number of iterations for the k-means algorithm. In addition, they can choose a distance measure (with Euclidean as default).

## 4.2 Consolidation

We use the cluster ensembles framework of Strehl and Ghosh [131] to combine the multiple clusterings into a single consolidated clustering. This popular technique aims at finding a consolidated clustering that maximizes the mutual information with a set of input clusterings. Mutual information is especially suited for our application because modelers are interested in the geographic regions where the input clusterings agree most on.

Since maximizing the mutual information is computationally prohibitive, we apply the three heuristics suggested in the cluster ensembles framework: the cluster-based similarity partitioning algorithm (CSPA), the hypergraph partitioning algorithm (HGPA), and the meta-clustering algorithm (MCLA). All heuristics first transform the set of input clusterings into a hypergraph. CSPA uses the relationships between objects expressed through the input clusterings to construct a measure of pairwise similarity. This measure can then be used with any similarity-based clustering algorithm (k-means in our case, see Section 4.1). HGPA performs a minimum cut operation to approximately maximize the mutual information. MCLA uses the hypergraph to identify and consolidate meta-clusters. Out of the consolidations resulting from the three heuristics, the one that has the highest average normalized mutual information (ANMI) with the input clusterings is chosen (see [131] for

further details).

While HGPA determines the optimum number of clusters automatically, the other two heuristics allow for controlling the final number of clusters. To provide modelers with the required flexibility in the consolidation process (AR3), our tool enables them to apply MCLA and CSPA for a varying number of final clusters. They can then choose to be presented with the best result in terms of information shared, or select any of the other consolidated clusterings that were created by the three heuristics.

## 4.3 Quantitative measures

To support the assessment of the consolidated clusterings (AR4) we compute three quantitative measures.

The first captures the information that a particular cluster in a consolidation result shares with the input clusterings. For its computation, we use the ANMI criterion from the cluster ensembles framework [131], but assume that the consolidated clustering only contains this particular cluster. We call this measure *marginal ANMI*. It allows modelers to identify clusters that should not be interpreted as geophysical processes.

We further compute the geospatial similarity as well as the feature similarity between consolidated clusters. These similarity measures are important criteria in the comparison of two data sets because they allow modelers to identify pairs of clusters that represent the same geophysical process. For the pairwise geospatial similarity, we compute the percentage of geographic overlap between clusters.

The pairwise feature similarity is determined as follows (assuming that multiple feature spaces were used to produce the input clusterings): first, we calculate the distance in each feature space between the centroids of two consolidated clusters; second, we normalize the separate distances, weigh and combine them. The weights for each feature space are assigned according to the number of input clusterings that were conducted in the respective feature space. Hence, the weights are implicitly provided by modelers since they create the set of input clusterings for the consolidation. Lastly, we convert this single distance measure into a similarity score.

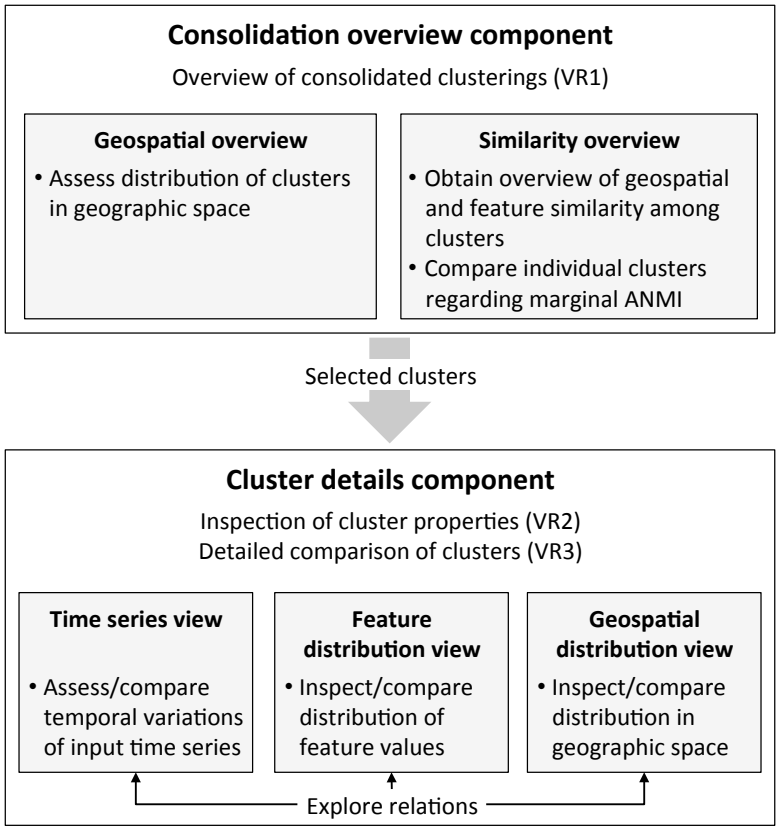



**Figure IV.2:** The components and views of our visual interface, and the visualization requirements and tasks they support. The consolidation overview component comprises two views which support modelers to obtain an overview of the consolidated clusterings. Based on this overview, modelers select clusters for detailed inspection and comparison in the cluster details component.

## 5 Interactive visual exploration and comparison

To meet the visualization requirements outlined in Section 3.3, our visual interface for exploration, interpretation, and comparison of the consolidated clusterings comprises two types of coupled visualization components (Figure IV.2). A *consolidation overview component* enables users to gain a basic understanding of relations among clusters in geographic space and in feature space, and to decide on subsequent analysis steps (VR1). Researchers select clusters in this component and pass them to a *cluster details component* where they can inspect the properties of a single cluster (VR2) but also compare multiple clusters in detail (VR3). Both components allow for visual queries to support the identified analysis tasks.

To establish a visual link, all clusters are color-coded consistently across the consolidation overview and cluster detail components. To this end, we use one of ColorBrewer's qualitative color schemes [52] as well as colors sampled from the CIELAB color space (see Guo et al. [35] for a suitable sam-

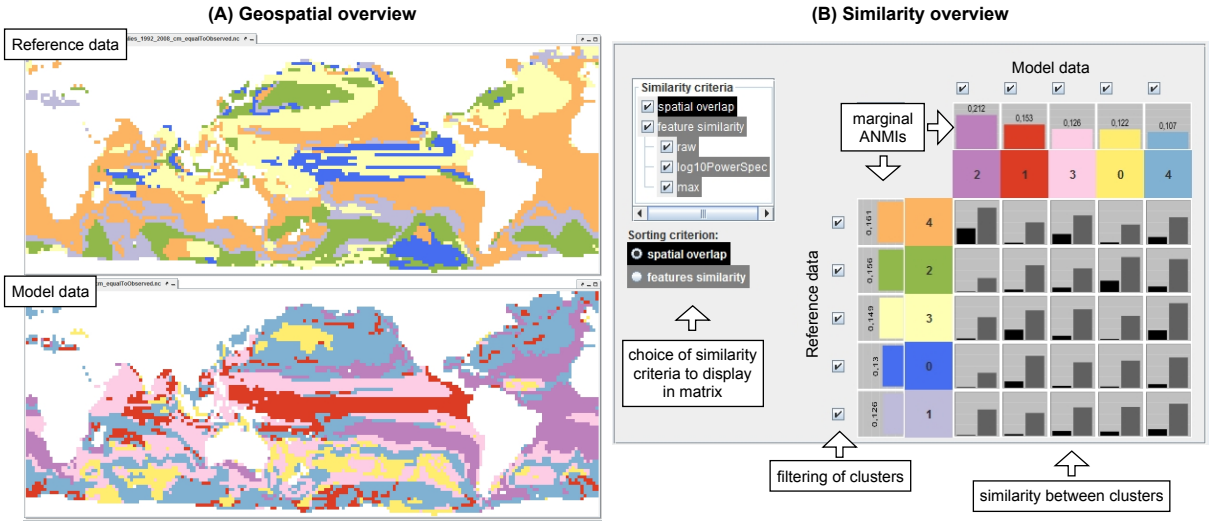



**Figure IV.3:** The consolidation overview component. Note that each cluster has its unique color since the cluster ensemble computes different clusters in both data sets. The consolidation overview supports scientists to focus on clusters that allow for a meaningful interpretation and to understand the relationships between clusters by comparing their geospatial and feature similarity.

pling strategy). We chose the ColorBrewer colors to provide users with carefully designed and easily distinguishable colors. If additional colors are required, we use the CIELAB samples. This strategy yields a sufficient number of distinguishable colors. In addition, users can change the colors manually to adjust the color coding according to their preference. Note that we assign a unique color to clusters in both data sets since the consolidation process computes different clusters in both data sets.

The tight integration of the analytical part (clustering and consolidation via cluster ensembles) allows scientists to change the configuration of the consolidation as well as the set of input clusterings. The former is done to improve the balance between quality and complexity of the consolidated clusterings (see AR3 in Section 3.3); the latter allows modelers to change the features considered in the analysis. They can make these changes at any time during visual exploration and study the resulting consolidated clusterings.

In the following, we explain the visual encoding and interactive capabilities of each visualization component, and how they contribute to the visualization requirements VR1–VR3.

## 5.1 Consolidation overview component

This component provides an overview of the two consolidation results (VR1) and, thus, acts as a starting point for the visual comparison process. It includes two views: a *geospatial overview* and a *similarity overview* (Figure IV.3).

### Geospatial overview

This view depicts the geospatial distribution of clusters in two maps; one for each consolidated clustering. The maps are juxtaposed to allow for visual comparison; cluster membership of geographic locations is encoded with a unique color for each cluster.

The geospatial overview allows users to observe a variety of patterns. For example, the maps in Figure IV.3 depict clusters that form quite coherent geospatial structures – e.g., the orange cluster in the model data – but also clusters with members that are more distributed over geographic space – e.g., the yellow cluster in the reference data. Notice also the different sizes of clusters as well as the horizontal structures along the equator.

Modelers interpret these patterns based on their domain knowledge and identify clusters that may represent geophysical processes. For example, the red cluster along the equator in the reference data probably represents a process in the tropics.

### Similarity overview

This view provides quantitative information about the clusters and their pairwise similarity (Figure IV.3) to support modelers to develop a first understanding of differences and similarities between model data and reference data. For a compact visual overview, the clusters are arranged in a matrix layout. The rows represent the clusters from model data; columns represent the clusters in the reference data. The clusters are ordered with respect to their marginal ANMI score. Each matrix cell contains a bar chart that depicts the geographic similarity and the feature similarity (see Section 4.3) between a pair of clusters. This enables modelers to quickly detect similar clusters.

The similarity overview also allows modelers to assess the clusters regarding the information they share with the input clusterings. For this purpose, we visualize the marginal ANMI (see Section 4.3) for each cluster as colored bars next to the cluster labels in the row and column headers. We arrange the ANMI bars in this way to facilitate comparison of clusters by judging position along a common scale. Since users were primarily interested in a relative comparison of clusters, the bars are scaled to the highest marginal ANMI score among all clusters. In addition, the bars are labeled with their ANMI values to help users judge the absolute amount of information shared with input clusterings.

### Linking and interaction

To support modelers in the assessment of relations among clusters, the similarity overview offers several filtering and sorting options.

In order to allow scientists to focus on a particular aspect of cluster similarity, we provide a checkbox tree (Figure IV.3) where they can select between displaying information on geospatial similarity, feature similarity, or both. They can further select the features to be considered in the computation of the feature similarity.

To obtain a better overview of similarities between clusters, modelers can also change the order of clusters in the matrix by mouse-clicking on a cluster label. When users click on a row label, the clusters in the columns are sorted in order of decreasing geospatial or feature similarity to the selected cluster, and vice versa. Users may switch the sorting criterion any time during analysis.

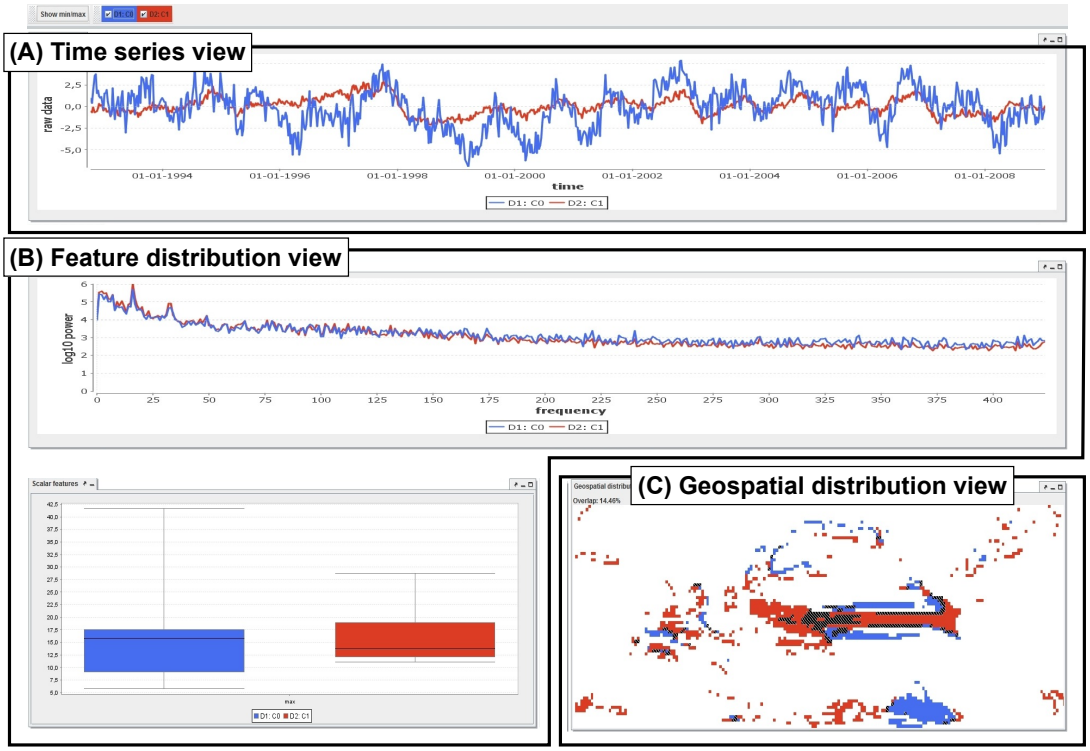
Lastly, checkboxes adjacent to the marginal ANMI bars allow modelers to visually filter clusters that they consider irrelevant or that share very little information with the input clusterings. Deselected clusters are greyed out in the similarity overview as well as in the geospatial overview.

An important functionality of the consolidation overview component is to enable users to choose and pass clusters to cluster details views for detailed inspection and comparison. To this end, modelers may either use the geospatial overview or the similarity view. In the geospatial overview, they may click on one of the maps to select a cluster and pass it to a cluster details view. In the similarity overview, modelers can click on a matrix cell to compare the two clusters associated with that cell in a new cluster details view.

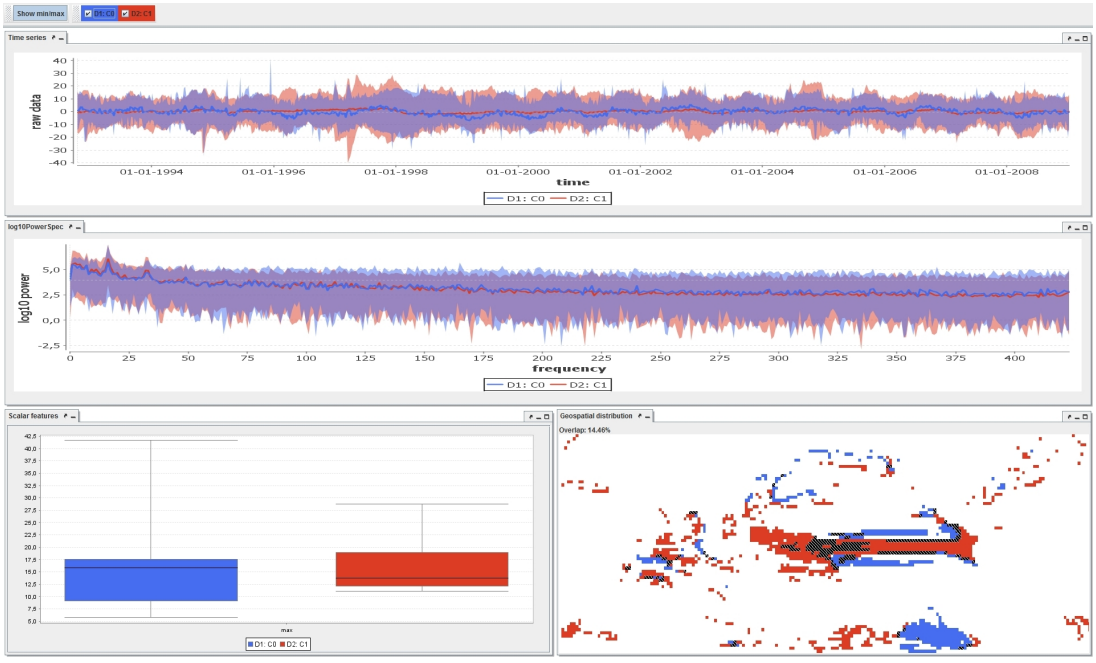
## 5.2 Cluster details component

This visualization component allows modelers to inspect the properties of a single cluster (VR2) to determine whether it represents a geophysical process. It also enables them to compare multiple clusters in detail (VR3) to study differences and similarity between clusters in model and reference data.

A cluster details component comprises three visualizations (Figure IV.4a). The *time series view* depicts the temporal signatures associated with clusters, the *feature distribution view* helps users to inspect and compare the distribution of feature values for each cluster, and a *geospatial distribution view* facilitates detailed inspection and comparison of clusters in geographic space. All three views are tightly-coupled and allow users to conduct visual queries to the consolidation results via linking




(a) The three views of a cluster details component.




(b) Cluster details component with semi-transparent minimum-maximum range ribbons in line charts.

**Figure IV.4:** The cluster details component.

and brushing.

#### Time series view

This view presents the time series of the cluster centroid – which is the average time series over all cluster members – in a line chart. To assess the range of temporal variations in clusters, modelers can choose to display a semi-transparent minimum-maximum ribbon around each cluster representative (Figure IV.4b). The colors of the semi-transparent ribbons from different clusters mix in areas of overlap. This allows users to visually compare the temporal variations of multiple clusters. The time series view also enables users to interactively change the time frame via zooming to focus on particular time periods of interest.

#### Feature distribution view

This view supports inspection and comparison of clusters regarding their distributions of feature values. Each feature is depicted in a separate visualization to reduce visual complexity. In accordance with the preference of our collaborators, we use line charts for features like the power spectrum of a temporal profile, and box-and-whisker plots for scalar features such as standard deviation. These two visualization types are appropriate for all features that our partners use for their analyses. To accommodate other types of features, this view can be easily extended to include additional visualizations such as star plots, 3D charts, or scatterplot.

The line charts are constructed in the same way as the time series view; cluster representatives are shown and optionally surrounded by a semi-transparent minimum-maximum ribbon (Figure IV.4b).

Box-and-whisker plots communicate comprehensive information about the distribution of feature values in clusters. Furthermore, juxtaposing box-and-whisker items from different clusters allows users to judge position along a common scale and, thus, facilitates comparison.

#### Geospatial distribution view

In this visualization, all clusters in the cluster details component are plotted in the same geographic space to facilitate detailed comparison of clusters from different consolidated clusterings. Areas where clusters overlap are highlighted to point researchers to similarities in geographic space (black areas in geospatial distribution views in Figure IV.4). This view also shows the total percentage of overlap between all selected clusters as quantitative information above the plotting area.

### Linking and interaction

The cluster details component provides several means of interaction to support inspection and comparison of clusters.

Apart from tooltips and zooming functionality, users can filter the clusters in this component using checkboxes at the top. This enables modelers to reduce visual complexity, for example, to focus on varying pairwise comparisons, to revisit the properties of a single cluster, or to assess how much a particular cluster contributes to the total geospatial overlap.

To allow modelers to explore relations between input time series, feature values, and geographic distribution of clusters, this visualization component offers three brushing mechanisms. First, researchers can brush a range of feature values in a box-and-whisker plot. The distributions of all cluster members that fall within the selected range are highlighted in the other plots in the cluster details component. Second, modelers may apply a vertical line brush in a line chart to determine all cluster members with values in the selected y-axis range at the specified x-axis index. Again, the corresponding distributions are highlighted in the other visualizations. Third, modelers can select either all overlapping or all non-overlapping geographic locations in the geospatial distribution view to study the distributions of features for these regions in the other views.

## 6 Application example: Ocean Model for Circulation and Tides

In this section, we explain how our approach supported the assessment of the Ocean Model for Circulation and Tides (OMCT) [134]. On the domain expert side of our collaboration in this particular example were two ocean modelers, one of them a leading OMCT expert and also co-author of this paper.

The OMCT simulates currents and tides of the global ocean and is used for removal of aliasing artifacts from observational data produced by the Gravity Recovery and Climate Experiment (GRACE) satellite mission [132]. GRACE data are used in the geosciences to gain valuable insight into a number of important processes on Earth, e.g., ice-mass changes, ocean tides, or Earth crust displacements associated with major earthquakes. To ensure a high quality of these widely used data, noise correction with models such as the OMCT is crucial.

The OMCT yields volumetric time series data, representing the ocean as 13 vertical depth layers of regular horizontal grids. Although the data comprise three geographic dimensions, an initial assessment of model output only requires to analyze the topmost layer, the sea surface, since most

mechanisms within the ocean manifest themselves in changes to sea surface heights.

The most important ocean processes to consider during the assessment of OMCT data are the western boundary currents (WBC) and the antarctic circumpolar current (ACC). These ocean currents cause high spatial and temporal variability of sea surface heights around Antarctica, South Africa and on the northwest-side of the Atlantic and Pacific Ocean basins.

The remainder of this section describes how our tool supported the assessment of a new version of the OMCT ($OMCT_{new}$) regarding its ability to depict WBC and ACC processes.

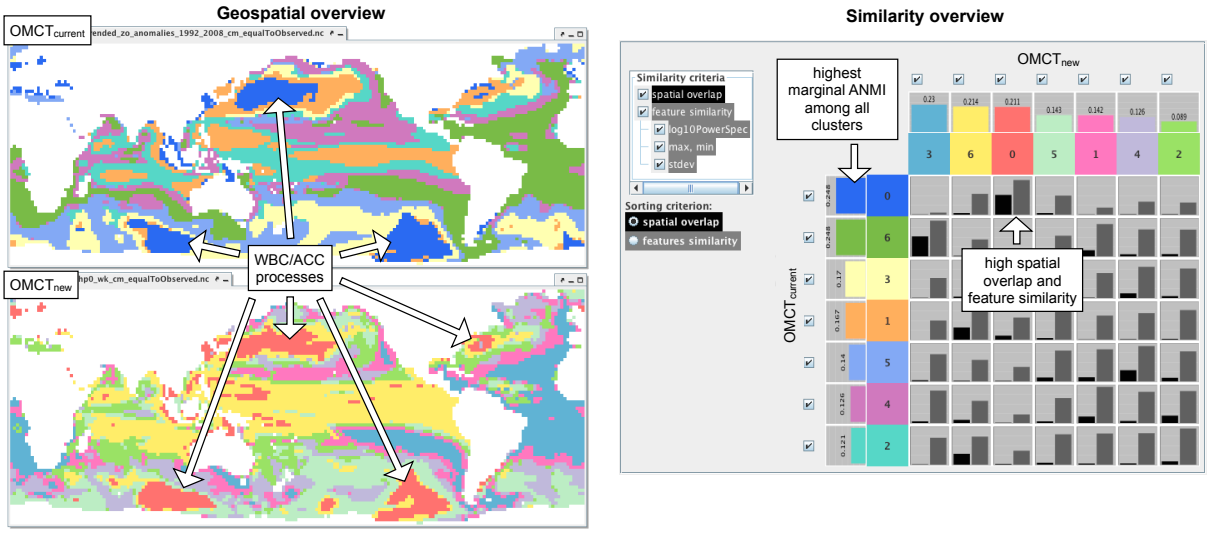
## 6.1 Results

To asses the $OMCT_{new}$, modelers compared sea-level anomalies simulated with this version with data produced by the current state-of-the-art OMCT version ($OMCT_{current}$). Each data set comprised approximately 9000 time series.

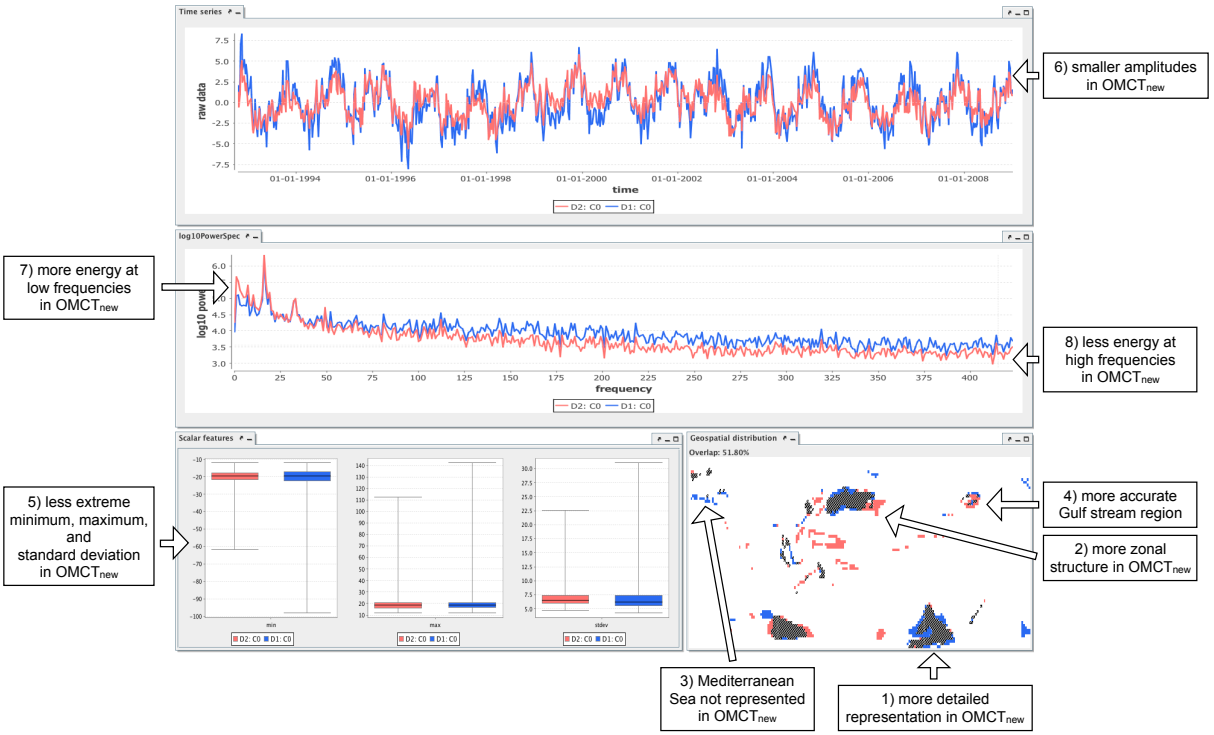
Applying our approach to the comparison of the two OMCT versions allowed for three important accomplishments: (1) the consolidation enabled modelers to capture the WBC and ACC processes in both data sets, (2) the visual interface permitted modelers to readily identify the clusters that represent these processes and to study relations between these clusters across data sets, and (3) detailed visual comparison helped them two determine that the $OMCT_{new}$ significantly improves the state-of-the-art OMCT version.

In the following, we provide more details about each of these three accomplishments.

**(1) Capturing WBC and ACC processes.** To describe WBC and ACC, modelers chose standard deviation, minimum and maximum, and the logarithmic power spectrum as features. These features describe the variation in a temporal profile from varying angles and in different granularities. Standard deviation summarizes the variability in a temporal profile in a single scalar value. The minimum-maximum feature is a vector of size two that provides information about the range of values that can be found in a temporal profile. The logarithmic power spectrum is a vector of size $n/2 - 1$ where $n$ is the length of a time series. It provides detailed information about the energy that a temporal profile exhibits at particular frequencies. Out of these three features, modelers previously could only use standard deviation to capture the variability of temporal profiles and to compare them in geographic space. Likewise, the power spectrum could only be used for detailed comparison of temporal profiles at a few geographic coordinates. Our visual analytics approach allows modelers to use all three features to compare the temporal profiles at all geographic coordinates.




**(a)** Consolidation overview showing consolidated clusterings for $OMCT_{current}$ (upper left) and $OMCT_{new}$ (lower left). The dark blue cluster and the red cluster represent western boundary currents and antarctic circumpolar currents. The similarity overview indicates that both clusters share a relatively high amount of information with the input clusterings and have high spatial overlap and feature similarity.




**(b)** Cluster details component comparing WBC/ACC clusters from $OMCT_{current}$ (dark blue) and $OMCT_{new}$ (red). $OMCT_{new}$ improves the simulation in various aspects. The red $OMCT_{new}$ cluster is a more accurate geographic representation of WBC/ACC processes (1-4), it also exhibits less extreme temporal behavior (5, 6), as well as an improved power spectrum (7, 8).

**Figure IV.5:** Comparison of a new version of the Ocean Model for Circulation and Tides (OMCT) – $OMCT_{new}$ – with the current state-of-the-art OMCT version – $OMCT_{current}$.

Our collaborators concluded that WBC and ACC processes can be differentiated into at least two but not more than twelve different types. Therefore, they chose k-means clustering for each of the three features with k ranging from two to twelve as input for the consolidation. This resulted in 33 input clusterings for each data set. Modelers also set the number of clusters for the consolidated clusterings to range from two to twelve. The consolidated clusterings with twelve clusters shared the most information with the sets of input clusterings. However, modelers determined with our tool that the results with only seven clusters had a much better balance between quality and complexity. In particular, the number of clusters was reduced by 40%, while the information shared with the input clusterings decreased by only 0.2% for $OMCT_{current}$ and 4% for $OMCT_{new}$.

**(2) Identifying WBC/ACC clusters in both data sets.** First, the geospatial overview in our tool allowed modelers to scan the consolidated clusterings for clusters that represent WBC and ACC processes (Figure IV.5a left). Based on their knowledge about the geospatial distribution of WBC/ACC and the geospatial shapes of the clusters, modelers quickly identified the dark blue cluster in the $OMCT_{current}$ data and the red cluster in the $OMCT_{new}$ as candidates.

Next, modelers were able to discern from the marginal ANMI scores in the similarity overview (Figure IV.5a right) that these two clusters share significant information with the input clusterings, and, hence, can be interpreted as geophysical processes. Of all clusters in the two data sets, the dark blue WBC/ACC cluster had the highest marginal ANMI, while the score for the red cluster was also relatively high. Modelers could also tell from the similarity overview that the two clusters have high geospatial and feature similarity and, thus, represent the same geophysical process.

**(3) Detailed comparison of model versions.** In developing a new model version, ocean modelers wanted to improve a number of aspects of the current state-of-the-art OMCT version ($OMCT_{current}$): (a) more detailed simulation of WBC and ACC regions, (b) improved simulation of WBC in the Gulf stream region, (c) smaller amplitudes in sea-surface heights over time, and (d) more energy for low frequencies and less energy for high frequencies in the power spectra of temporal profiles.

Exploring the two WBC/ACC clusters in a cluster details component (Figure IV.5b) allowed for assessing all of these aspects. The geospatial distribution view enabled modelers to notice four geographic characteristics that have improved with the new model version: (1) the region west of South Africa (lower right overlap) is represented in much more detail, (2) the WBC region in the northwest Pacific (top middle overlap region) is simulated in a more zonal structure, (3) the Mediterranean Sea is not represented in the red $OMCT_{new}$ cluster, and (4) the Gulf stream region is also represented

more accurately in the $OMCT_{new}$ cluster. In sum, modelers concluded from the geospatial distribution view that the new model version provides a more accurate geographic representation of WBC and ACC processes.

In addition, the time series view and the feature distribution view show that the red $OMCT_{new}$ cluster exhibits less extreme temporal behavior with smaller amplitudes. Lastly, a comparison of the logarithmic power spectra (Figure IV.5b) reveals that the final objective has also been met. In comparison to the $OMCT_{current}$ cluster, the $OMCT_{new}$ cluster has more energy at low frequencies and less energy in higher frequencies.

All the above findings constitute aspects of potential improvement in the new version of the OMCT – aspects that our partners could readily study with the help of our approach. After additional statistical analyses to corroborate the increase in quality, the $OMCT_{new}$ became the new state-of-the-art OMCT version.

## 6.2 User feedback

Our partners consider our approach a valuable complement to their existing tools and routine for four primary reasons: (1) they are not limited any more to single statistical measures for the detection of geophysical processes, instead, they have access to a range of features of temporal behavior, (2) the combination of multiple clusterings and cluster ensembles improves the detection of geophysical processes because multiple features of temporal behavior are considered simultaneously, (3) our interactive tool enables modelers to obtain a more complete picture about differences and similarities between model and reference data, and (4) it has great potential for speeding up the model development process because it supports a quick initial assessment of new model versions.

Our partners also value the flexibility of our approach. Our tool can be extended to include any feature that describes the temporal behavior represented in a temporal profile and, hence, allows scientists to study any geophysical process that may be of relevance to the assessment. Modelers are also highly flexible regarding the visual exploration. Although the consolidation overview component provides them with important information that guides subsequent analysis steps, they can always decide to make an educated guess and readily study any potentially interesting aspect. Before they had our tool, such an educated guess was not feasible because it involved time-consuming scripting and plotting procedures, a problem that also exists in related domains such as climate research [113, 135]. The means of interaction provided in our tool effectively remove this hurdle.

### 6.3 Discussion

Although our concept provides significant benefits to modelers in the assessment of ocean models, several issues need to be discussed.

First, the quality of the input clusterings determines the quality of the consolidated clusterings. Therefore, the features chosen for the clusterings must capture geophysical processes. If not, the consolidation will not yield meaningful clusters. However, since our approach was developed with and for expert users, one can assume that the features chosen will be appropriate for the respective analysis task.

Second, the computational complexity of the consolidation process is quadratic in the number of temporal profiles in a data set. Although this can be addressed in future work, it has not been a major issue in our application for two reasons. (1) A single simulation run of an ocean model typically takes several days if not weeks; in this context, the required time for the consolidation process is negligible. (2) In the opinion of modelers, the benefits of our approach outweigh the downside of the high computational complexity. Depending on the hardware available, the automated analysis part of our approach is applicable to models in the range of 100K grid points.

Another point worth mentioning is that the time series view is currently limited to a reasonable number of time steps in the temporal profiles. In practice, however, this was no issue because ocean modelers typically study weekly, monthly, or even seasonal averages. This approach is also applied by climate scientists (as described by Poco et al. [113]) and significantly reduces the number of time steps per year.

## 7 Conclusion and future work

In this article, we presented a visual analytics concept that addresses a crucial part in ocean modeling: the comparison of model output with reference data. This concept was developed in close collaboration with ocean modelers, which allowed us to identify the primary challenges: the drastic aggregation that had to be performed and the high degree of subjectivity in the comparison process. To address these challenges we integrate clustering ensembles and interactive visual analysis into a tightly-coupled system. This approach is based on a comprehensive task and requirement analysis.

We have shown that the combination of data mining and interactive visual analysis can be of high value to the assessment of ocean models. We could also observe in our collaboration that the promising results of our work have led to increasing acceptance of visual analytics in the ocean

modeling community.

To further enhance our approach, we identified several major areas of future work. The next step is to conduct an in-depth user study to further corroborate the promising results we were able to achieve so far. We also want to extend our concept to very high-resolution ocean models, which requires improving the scalability of the visualization components as well as reducing the time and storage complexity of the consolidation. Therefore, we would like to investigate the applicability of distributed computing and GPU processing to our concept. We also plan on working on efficient algorithms to further speed up the consolidation and to better adapt to the characteristics of ocean model data. Currently, our concept supports the analysis of sea surface heights or any other meaningful two-dimensional layer. To apply our approach to all three geospatial dimensions, we will, again, have to address the scalability of the automated analysis, but in addition, identify and meet the visualization requirements that come with the third spatial dimension. We would also like to extend our approach from a two-way comparison to a three-way comparison. This would support an even more comprehensive assessment. A three-way comparison, however, also introduces additional challenges for visual analytics. Finally, our vision is to incorporate other types of geoscientific simulation models, beginning with climate models since their characteristics are somewhat similar to ocean models.

## Acknowledgements

The authors thank Joachim Fohringer, Ralf Friedeman, and Alexander Bobach for their help in implementing the prototype. This study was partially supported by the German Federal Ministry for Education and Research (BMBF) via the Potsdam Research Cluster for Georisk Analysis, Environmental Change and Sustainability (PROGRESS) (grant number 03IS2191A).

# Chapter V

# Visual analytics for correlation-based comparison of time series ensembles

**Published as:** P. Köthur, C. Witt, M. Sips, N. Marwan, S. Schinkel, and D. Dransch. Visual Analytics for Correlation-Based Comparison of Time Series Ensembles. *Computer Graphics Forum*, 34(3):411–420, doi:10.1111/cgf.12653, 2015.

**Abstract.** An established approach to studying interrelations between two non-stationary time series is to compute the 'windowed' cross-correlation (WCC). The time series are divided into intervals and the cross-correlation between corresponding intervals is calculated. The outcome is a matrix that describes the correlation between two time series for different intervals and varying time lags. This important technique can only be used to compare two single time series. However, many applications require the comparison of ensembles of time series. Therefore, we propose a visual analytics approach that extends the WCC to support a correlation-based comparison of two ensembles of time series. We compute the pairwise WCC between all time series from the two ensembles, which results in hundreds of thousands of WCC matrices. Statistical measures are used to derive a concise description of the time-varying correlations between the ensembles as well as the uncertainty of the correlation values. We further introduce a visually scalable overview visualization of the computed correlation and uncertainty information. These components are combined with multiple linked views into a visual analytics system to support configuration of the WCC as well as detailed analysis of correlation patterns between two ensembles. Two use cases from very different domains, cognitive science and paleoclimatology, demonstrate the utility of our approach.

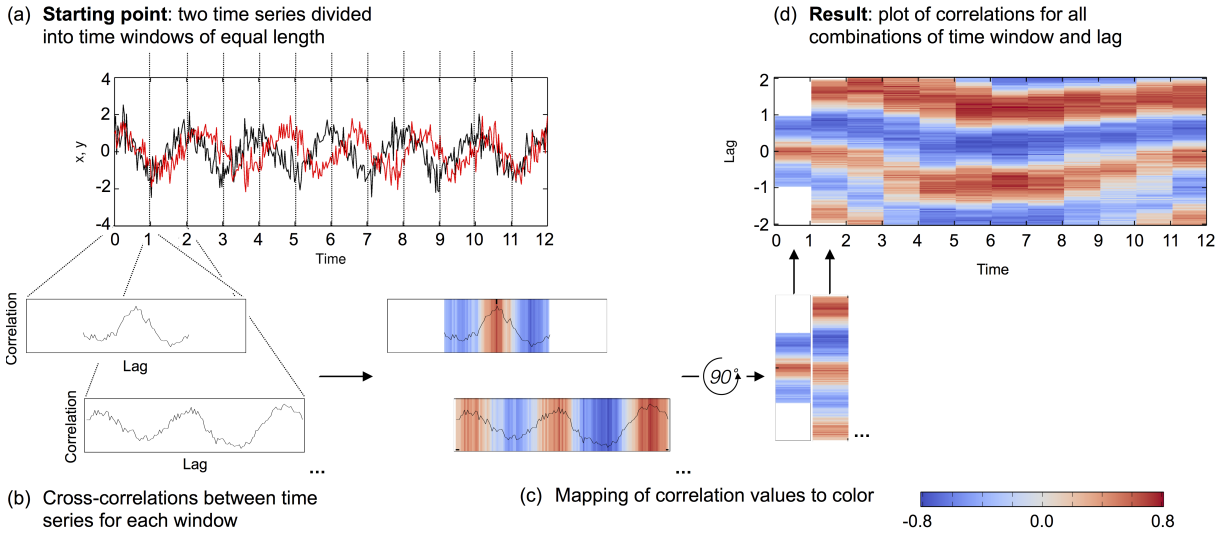



**Figure V.1:** Computation and basic plot of the windowed cross-correlation between two time series. The time series are divided into windows of equal length (a) and the cross-correlation between the time series for corresponding intervals is calculated (b). The resulting matrix of correlations can be visualized by mapping correlation to color (c). The matrix plot shows that one time series exhibits a time-varying phase difference, visible by the change of the lagged correlation over time (d).

# 1 Introduction

Time series are analyzed in many scientific disciplines. An essential analysis task is the correlation-based comparison of time series. It allows for studying important phenomena such as climate impacts on the spreading of hazardous infectious diseases [53]. A powerful technique for this task is the windowed cross-correlation (WCC) [4, 20]. It addresses two frequent problems: (1) there can be a time lag between the temporal behavior in different time series [49], and (2) the time series can be non-stationary and, thus, their correlation can change over time [160]. The WCC solves these problems for the comparison of two time series as follows: It divides the time series into intervals of equal length, called windows (Fig. V.1a). Then, the cross-correlation (CC) between corresponding windows is calculated (Fig. V.1b). The CC shifts two windows relative to each other and computes the correlations between them for a range of temporal offsets (lags) [96]. The outcome of the WCC is a two-dimensional matrix of correlations between two time series. The columns represent the position of the time windows, the rows the different lags. A static plot of this matrix, in which correlation values are mapped to color (Fig. V.1c and V.1d), is typically used by scientists to investigate the time-varying correlation between two time series [98].

Although powerful, the WCC only address the comparison of two individual time series. How-

ever, many scientific disciplines such as climate modeling [21], chemistry [126], or brain research [13] study ensembles of time series (sets of time series produced, e.g., through Monte Carlo simulation or repeated measurements). Therefore, we extend the WCC to a visual analytics solution that supports a correlation-based comparison of two ensembles of time series. Our approach combines semiautomatic statistical analysis with interactive visual exploration. The computational part enables users to calculate the pairwise WCC between all time series from the two ensembles. Since a single ensemble may easily comprise hundreds of time series, this yields hundreds of thousands or even millions of WCC matrices. Each matrix depicts the same combinations of time window and lag; only the correlation values differ between matrices. To support visual exploration, we use statistical measures to describe the resulting distribution of correlation values at each combination of time window and lag. The outcome of the statistical analysis is presented in a visually scalable overview visualization of the WCC. It depicts the sign, magnitude and uncertainty of the time-varying correlations between two ensembles. WCC computation and overview visualization are combined with multiple linked views into a visual analytics system. It supports flexible configuration of the WCC and detailed analysis of correlation patterns as well as relating these patterns to the input ensemble data.

This approach is based on a thorough task- and requirement analysis. After an overview of related work (Section 2), we will provide more detail about the identified requirements (Section 3). To meet these requirements, we had to face a number of visualization challenges, in particular, plotting and exploring more than a thousand time series with up to 10k observations each, visualizing large WCC matrices, as well as visualization and exploration of uncertainty information. Section 4 will describe the individual components of our concept and how they address the requirements and visualization challenges. We further provide two use cases from very different domains to demonstrate the significance of our approach (Section 5): (1) detection of event-related potentials in electroencephalography (EEG) measurements and (2) comparison of paleoclimate time series ensembles derived from stalagmites. Finally, we summarize our results and suggest areas of future work (Section 6).

## 2 Related work

Since we have covered studies concerning the computation and application of the windowed cross-correlation in the introduction, we will focus this section on related work from the visualization community.

While many works support visualization and visual exploration of time series data [2], only a limited number of systems focus on the visual analysis of ensemble data [73]. These approaches provide valuable solutions for applications such as weather forecasting [115, 120], finding potential indicators of climate change [74], car engine optimization [100], or development of power train systems [112].

We address the comparison of time series ensembles, in particular, the detection of time-varying correlations between two ensembles. For exploration of the ensemble data, we took inspiration from works that use binning to visualize time series with thousands of observations [17, 31], and approaches that map line density to opacity to visualize large sets of time series at interactive frame rates [103, 109]. To analyze the uncertainty of correlations between two ensembles, we studied various guidelines for encoding uncertainty information [93, 95, 161]. We were especially inspired by approaches that use statistical moments to visualize the uncertainty in distributions of numerical values [26, 59, 68, 110, 114, 115]. Since we must cope with a large number of distributions of correlation values, we turn to matrix visualization, which allows for displaying massive data in a compact visual overview [152]. We specifically build on a matrix visualization technique called Hinton diagram [57, 58]. This technique is used in network analysis to visualize, e.g., network weights or activations of units in a network [22, 87]. We extend the Hinton diagram to a matrix that supports hierarchical aggregation and semantic zooming [40, 41] to depict the time-varying correlations between two ensembles as well as their uncertainty.

## 3 Design requirements

The visual analytics approach presented in this article is the result of a close collaboration with two experts in time series analysis, both co-authors of this paper. A user- and task-centered approach [37] that involved frequent meetings and discussions allowed us to elicit the following design requirements for the comparison of two time series ensembles via WCC:

DR1 *Allow for flexible configuration of windowed cross-correlation (WCC) computation*

Scientists typically focus on particular aspects of correlations during the comparison of time series via WCC, such as short-, medium-, or long-term temporal variations of correlation. Each of these aspects requires a different set of parameters for the WCC computation. Furthermore, scientists sometimes want to focus the comparison on specific subsets of the ensembles. Both – analysis focus and subsets of interests – may change in the comparison process. Researchers

must therefore be able to:

- Choose between comparing entire ensembles or user-specified subsets.
- Modify the parameters and recompute the WCC.
- Browse and easily access WCC results previously computed in the analysis process.

DR2 *Provide measures that capture the uncertainty of computed correlation values*

Our collaborating scientists consider the spread of a distribution of correlation values as uncertainty. They require quantitative information about this spread for each combination of time window and lag in the WCC. A common approach is to calculate statistical moments that quantify the central tendency and the degree of dispersion. These measure can then be used in (interactive) visualizations to further facilitate the assessment of uncertainties. Therefore, a visual analytics approach must:

- Provide measures of central tendency.
- Provide measures of dispersion.
- Provide confidence intervals.

DR3 *Provide overview of correlations and their uncertainty*

Scientists want to detect the predominant patterns in the correlations between two time series ensembles, especially patterns of temporal variation and patterns of uncertainty of correlation values. This requirement is associated with the following tasks:

- Obtain overview of correlation values for all combinations of time window and lag.
- Obtain overview of the uncertainty of correlation values for all combinations of time window and lag.

DR4 *Support exploration of correlations and their uncertainty*

After detecting the predominant patterns of correlation and uncertainty, researchers want to gain a detailed understanding of the time-varying correlation between two ensembles as well as the reliability of the results. To this end, they need to:

- Inspect in detail the temporal variation of correlations and the variations in uncertainty.
- Assess the statistical significance of correlation values.

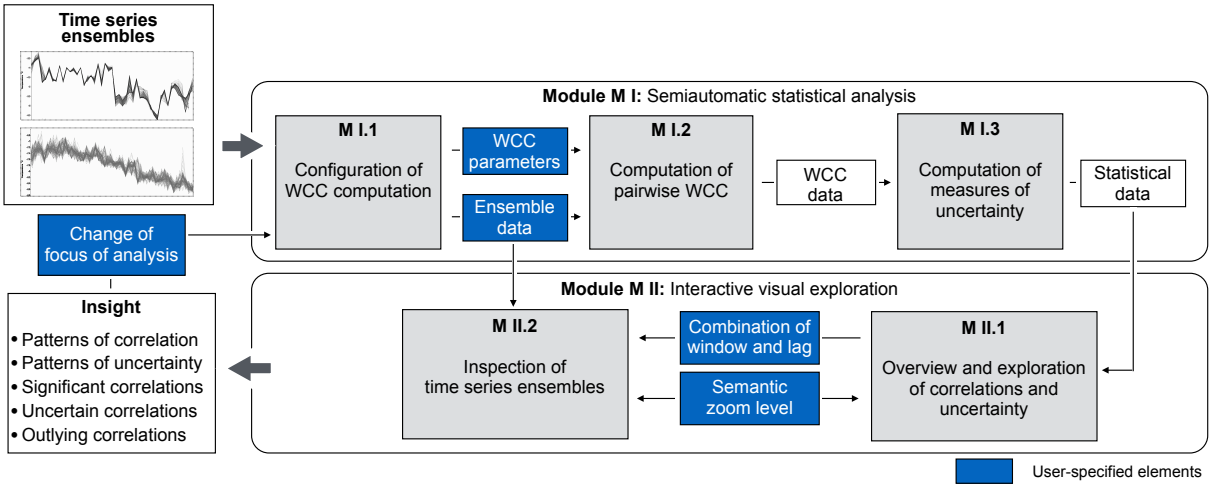



**Figure V.2:** Our visual analytics concept. It comprises computation and statistical analysis of windowed cross-correlations (WCC) between two time series ensembles (module M I) and interactive visual exploration of the resulting time-varying correlations and their uncertainty (module M II).

– Examine the distribution of correlation values for individual combinations of time window and lag.

DR5 *Support inspection of the two time series ensembles*

To interpret the results of the WCC computation and to gain a better understanding of the time-varying correlation, scientists have to inspect and compare the ensemble data that went into the calculation. In particular, they have to:

– Obtain an overview of the time series in the two ensembles.
– Inspect and compare the distribution of time series in the ensembles.

# 4 Visual analytics approach

To meet the identified requirements, our concept combines two modules: semiautomatic statistical analysis (module M I) and interactive visual exploration[1] (module M II). Each module comprises several components (Figure V.2).

## 4.1 Linking between modules and components

On a module level, users choose the ensemble data and parametrize the WCC calculation in module M I, which then computes the data for visual exploration in module M II. The insight researchers

[1] A supplementary demo video of the prototype illustrating the visual exploration part can be found at http://doi.org/10.2312/GFZ.1.5.2016.001

gain through interactive visual exploration may change the focus of analysis, which, in turn, may prompt users to go back and modify the WCC parameters or even the input data.

On the component level, M I.1 passes the user-selected ensemble data and the parameters for the WCC computation to M I.2. The latter calculates the pairwise WCC between the time series from both ensembles and forwards the resulting WCC data to M I.3. Component M I.3 computes statistical measures to summarize the correlations and their uncertainty. These data are passed to component M II.1 for visual exploration. This component provides researchers with an overview of the correlation results. It also allows users to interactively filter combinations of time window and lag that meet specified criteria regarding correlation values and uncertainty. The underlying distribution of correlation values for selected combinations of time window and lag can also be explored in M II.1, while the corresponding time series data can be studied in component in M II.2. Both components, M II.1 and M II.2, support hierarchical aggregation and semantic zooming to cope with the limited screen space.

The remainder of this section will describe the two main modules of our concept as well as their respective components and how they address the design requirements.

## 4.2 Module M I: Semiautomatic statistical analysis

This part comprises components M I.1, M I.2, and M I.3 which cover design requirements DR1 and DR2.

### Component M I.1: Configuration of windowed cross-correlation (WCC) computation

DR1 – *Allow for flexible configuration of WCC computation.*

Component M I.1 allows scientists to choose the two ensembles for comparison and, if need be, to further focus on particular time series from the two ensembles. Next, researchers use M I.1 to set the parameters for the WCC: window size, overlap between consecutive windows, and lag range. Scientists do not necessarily have to initiate a new computation. For every compared pair of ensembles, we store previous WCC configurations and results in a database. Scientists can use this component to browse through and to revisit them. When users decide to initiate a new computation, the ensemble data and the WCC parameters are passed to component M I.2.

#### Component M I.2: Computation of pairwise of windowed cross-correlation (WCC)

This component performs the actual WCC computation. Let $M$ and $N$ be the two sets (ensembles) of time series. All time series in $M$ and $N$ must be equal regarding the number of observations, the timestamps of the observations, and the length of intervals between observations. Note that M I.2 could also be extended to include correlation analysis techniques for irregularly sampled time series, such as kernel based correlation estimation [116]. We compute the WCC for all pairs of time series in $M \times N$ (see [20] for details regarding the computation of the WCC). As a result, we obtain $k = |M \times N|$ WCC matrices. These data are passed to component M I.3.

#### Component M I.3: Computation of measures of uncertainty

DR2 – *Provide measures that capture the uncertainty of computed correlation values.*

Since the same parameters were used for all WCC computations, each matrix depicts the same combinations of time window and lag; only the correlation values differ between matrices. Hence, we obtain a distribution of correlations for each window-lag combination. We use descriptive statistics for analyzing the uncertainty in these distributions [143]. This component calculates the mean and median correlation as measures of central tendency for each window-lag combination, and the standard deviation and interquartile range as measures of dispersion. It also determines the confidence interval of the correlations for a user-specified $p$-value with a $t$-test [143]. This information is then used to calculate the percentage of statistically significant correlations at each combination of window and lag. Note that M I.3 is not limited to this set of statistical measures. If need be, our concept allows for incorporating additional measures.

### 4.3 Module M II: Interactive visual exploration

This module is composed of component M II.1, which covers design requirement DR3 as well as DR4, and component M II.2, , which addresses design requirement DR5.

#### Component M II.1: Overview and exploration of correlations and their uncertainty

DR3 – *Provide overview of correlations and their uncertainty.*

To address this requirement, we extend the basic WCC matrix plot introduced in Figure V.1 to an interactive and visually scalable visualization of the WCC and its uncertainty. Our technique is inspired by Hinton diagrams [57,58]. We map each combination of time window and lag to a square

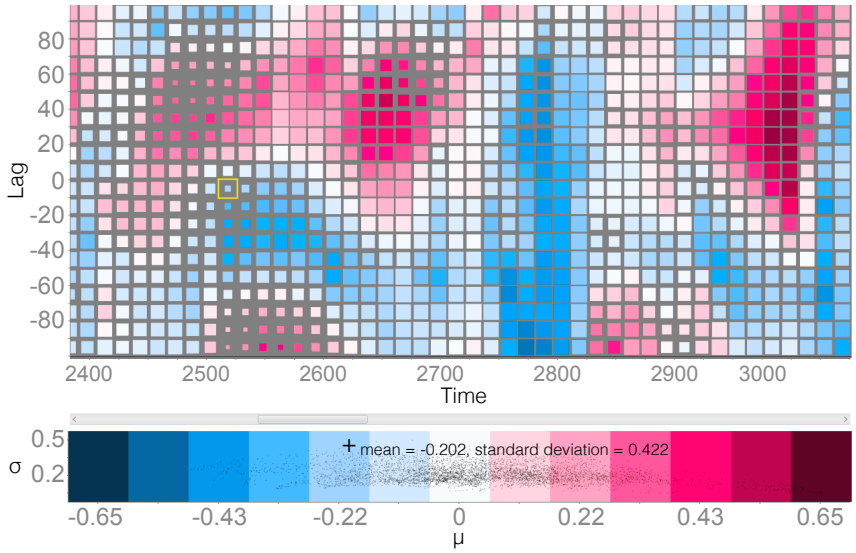



**Figure V.3:** Correlations view (top) and color legend with integrated scatter plot (bottom) of component M II.1. The color of the squares encode the mean or median correlations of each combination of time window (x-axis) and lag (y-axis). The size inversely depicts the uncertainty of the underlying distribution. Thus, small squares represent high uncertainty. When users select a window-lag combination it is highlighted in the correlations view (yellow square) and in the scatter plot (labeled cross).

in a matrix (Figure V.3). The color of the square denotes the mean of the underlying distribution of correlation values. We use a diverging color scale to differentiate positive and negative correlations of different magnitude. It is centered around zero and its range is determined by the highest absolute correlation value in the data. The square's size inversely encodes the standard deviation of the distribution. Hence, the higher the uncertainty (or variation), the smaller the square. For more robustness against outliers, users can alternatively choose to display the median and interquartile range in the correlations view. The statistical measures for this mapping are provided by component M I.3.

The texture that results from arranging the squares on the *XY* plane facilitates preattentive processing of patterns. This view enables scientists to differentiate strong correlations (saturated colors) from weak correlations (light colors) and uncertain values (small squares) from rather certain correlations (large squares). Note that color and size in this encoding potentially interact [32]. In our scenario, however, scientists are particularly interested in strong correlations that relatively certain. These represent the most reliable indicators of meaningful correlations between two ensembles. Since this information is emphasized in our design through large squares of intense color, there is little chance of misinterpretation.

Although Hinton diagrams are a powerful technique, they do not scale well to large WCC matrices because the squares encoding the numerical values require significant screen space. We

experimented with mapping uncertainty to transparency instead of size as a more space-efficient encoding. However, this alternative hindered the interpretation of correlation values. Therefore, we chose to extend the Hinton technique to support semantic zooming [41]. We bin the matrix whenever the screen space does not suffice to display all squares. The statistical information conveyed by the squares in each bin is summarized in a single glyph. We operate on the statistical information encoded in the input squares instead of the underlying distributions of correlations because it allows for on-the-fly construction of the glyphs and, hence, zooming of the matrix in real time. The glyph is composed of three nested squares (Figure V.4). The color of the middle square encodes the median correlation of the input squares; its size represents the median uncertainty. A transparent outer square denotes the uncertainty of the least uncertain input square. A transparent inner square shows the uncertainty of the most uncertain input square. We chose this design for three reasons: (1) it preserves the visual encoding of the input squares, (2) it conveys information about the range of uncertainty among the aggregated window-lag combinations, and (3) the glyphs are easily distinguishable from the squares of the non-aggregated representation, which signals to users that they are looking at visual aggregates. The binning of the matrix depends on the available screen space and the minimum sizes of the squares and glyphs. When users change the zoom level, the new binning as well as the input squares for each bin are determined, and the glyphs are adjusted accordingly.

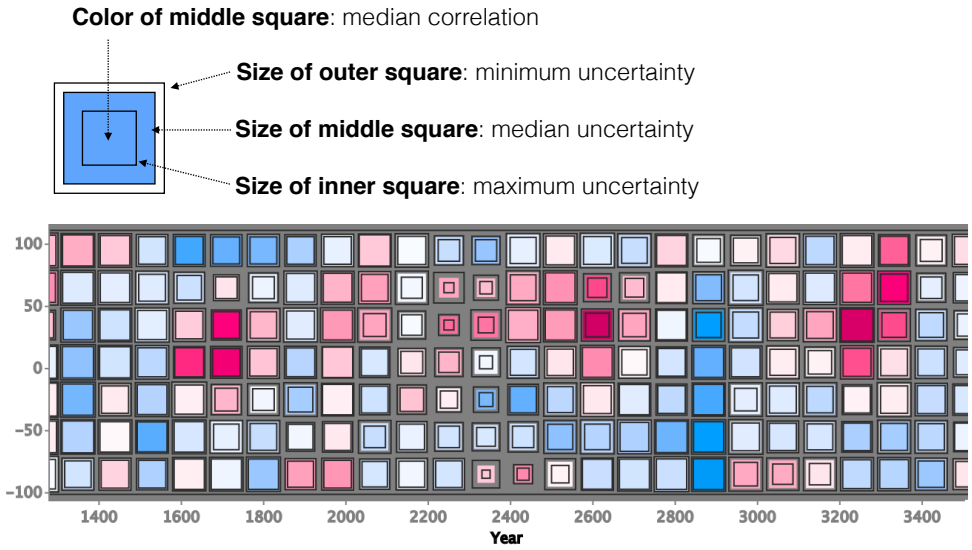



**Figure V.4:** Glyph encoding the results of aggregating the distributions of correlations from multiple window-lag combinations (top) and a zoomed-out version of the correlations view using the glyph (bottom). 'Uncertainty' denotes either standard deviation or interquartile range of correlation values. Small squares signal high uncertainty and vice versa.

To enable scientists to obtain an overview of potential relationships between correlation values and uncertainty across all window-lag combinations, we integrate a scatter plot into the color legend of the correlations view (Figure V.3). It shows the mean or median correlation values plotted against the uncertainty. Furthermore, the correlations view and the scatter plot are linked to enable exact quantitative assessment. When users mouse-point at any window-lag combination in the matrix, the exact correlation and uncertainty values are displayed in the scatter plot.

To present scientists with additional information about the uncertainty of correlation values, our tool provides an alternative view of the WCC matrix. This view maps the fraction of statistically significant correlations at each combination of time window and lag to color in a pixel display (see Figure V.10 for a view of the fraction of positive correlations).

DR4 – *Support exploration of correlations and their uncertainty.*

Besides a semantic zoom for detailed inspection of correlations and uncertainties, component M II.1 provides two additional mechanisms to meet requirement DR4.

The first mechanism is interactive filtering via range sliders. Scientists can use these sliders to gray out window-lag combinations in the correlations view that do not meet specified quantitative criteria. In particular, we offer the measures from component M I.3 for filtering: mean and median correlation, standard deviation, interquartile range, and the fraction of significant (negative/positive/total) correlations at the window-lag combinations. The correlations view adjusts in real time, which allows scientists to explore the variation of correlations and their uncertainty, as well as the statistical significance of the correlations.

To enable scientists to explore the distributions of correlation values that are represented by each square in the correlations view, we provide an on-demand histogram (Figure V.5). It allows researchers to examine properties of the underlying distributions, e.g., modality or symmetry. Furthermore, gray areas in the histogram mark the statistically significant portions of the distribution.

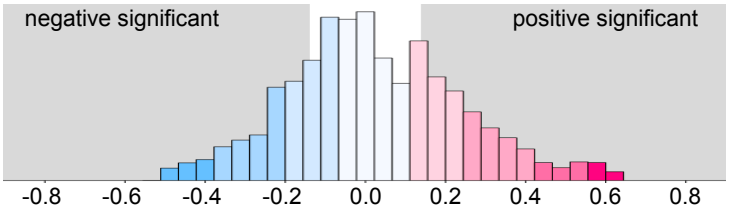



**Figure V.5:** On-demand histogram of correlation values for a combination of time window and lag. The gray areas in the background mark statistically significant correlations.

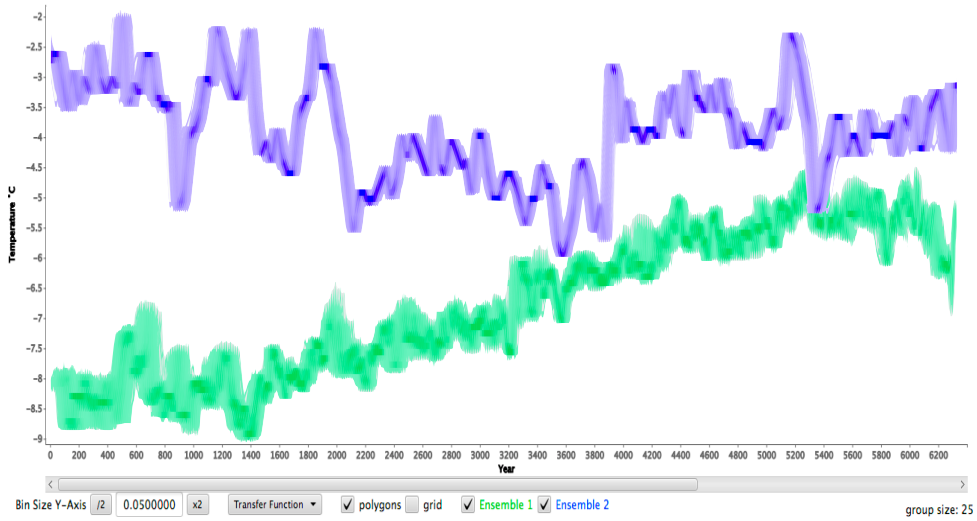



**Figure V.6:** Line chart showing both ensembles (component M II.2). Line density is mapped to opacity to depict the uncertainty in the ensembles. Dense regions in each ensemble are mapped to high opacity, sparse regions are more transparent to indicate less agreement among ensemble members.

### Component M II.2: Visual inspection and comparison of the two time series ensembles

DR5 – *Support inspection of the two time series ensembles.*

Component M II.2 visualizes the two time series ensembles that went into the WCC calculation (Figure V.6). Both ensembles are plotted in the same line chart with a unique color assigned to each ensemble. This facilitates inspection and comparison regarding trends, amplitude, scale, and range – valuable information that helps scientists to interpret the WCC results. Users can further choose between displaying the original time series data or normalized versions. The latter enables scientists to compare the two ensembles on the same scaling level while the former accentuates differences in scaling level.

To make M II.2 visually scalable, we had to address two issues: (1) visualization of time series with thousands of data points, and (2) depicting hundreds or thousands of time series in the same plot.

We use a binning approach in combination with semantic zooming to address the first issue. In particular, we divide the time series into intervals of equal length (bins) and calculate the mean of each interval. The resulting averaged time series are then shown in the line chart. Semantic zooming and panning enables scientists to inspect the binned time series in more detail. After each zooming action the bin size (level of aggregation) is adjusted automatically to match the number of pixels available to display the selected time range.

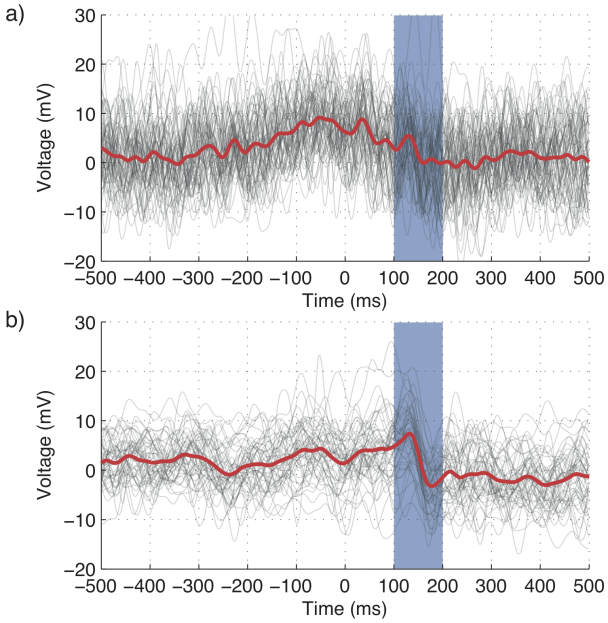



**Figure V.7:** Electroencephalogram (EEG) ensembles of two subjects (a) and (b). The red lines represent the average of the respective trial. After a visual stimulus at time $t = 0$ ms, an evoked electrophysiological activity around 170 ms (marked by the blue shading) can be clearly observed for subject (b), but only suspected for subject (a).

To address the second scalability issue, we map the line density of each ensemble to opacity [74, 103, 109]. Dense regions signal high agreement among ensemble members and are mapped to high opacity; sparse regions are more transparent to indicate less agreement among ensemble members. This provides users with an overview of the distribution of time series within the ensembles. If need be, check-boxes allow scientists to display, and therefore focus on, only one of the two ensembles.

## 5 Use cases

We demonstrate the utility of our concept with two applications from the fields of cognitive science and paleoclimatology. In both fields, ensembles of time series are frequently compared. However, the standard procedure is to compare only the mean or median time series of the ensembles. Our visual analytics approach enabled us to perform a correlation-based comparison of entire ensembles.

### 5.1 Interpersonal detection of event-related potentials

In cognitive science, the brain activity is studied by measuring the electroencephalogram (EEG) at the human scalp. In experiments, stimuli are presented to subjects and the potential changes in measured brain activity, called event-related potential (ERP), are investigated [92]. Due to many

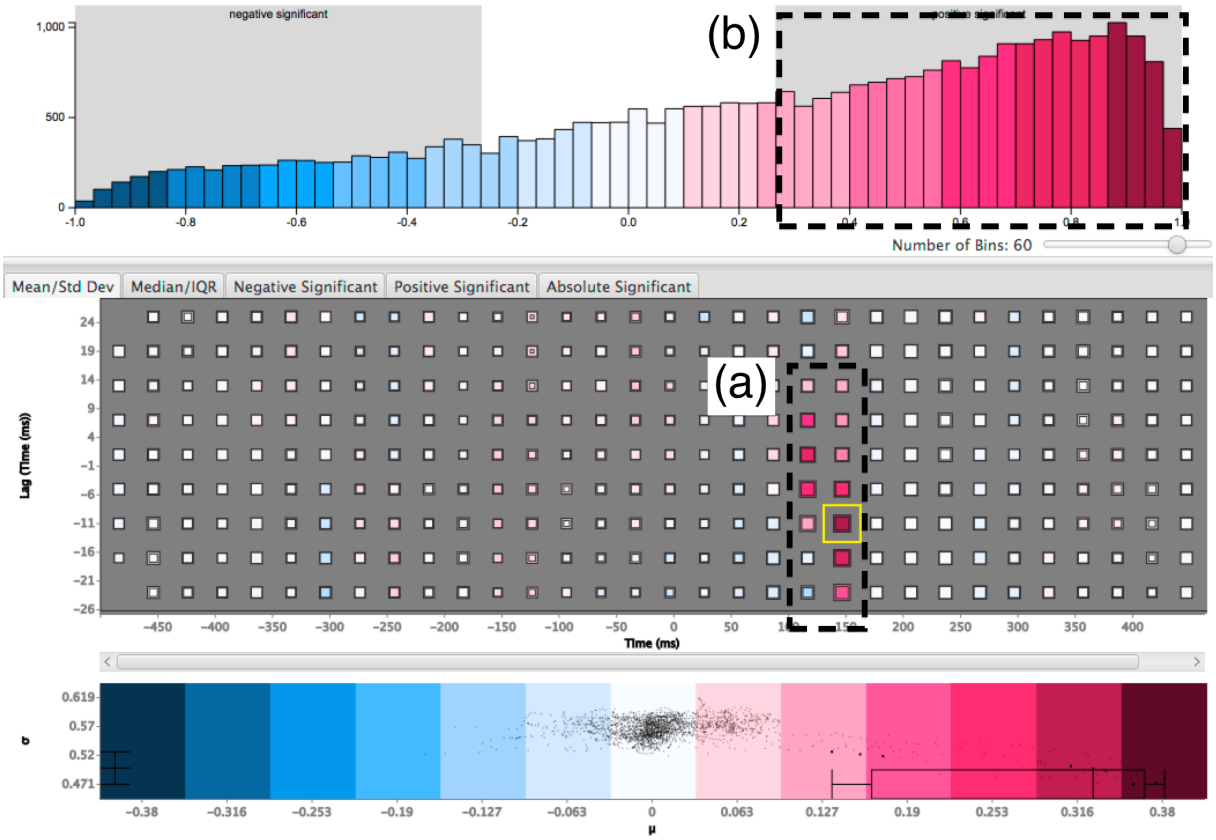



**Figure V.8:** Correlations between EEG ensembles of two subjects. In both subjects, the presented facial stimulus causes the same electrophysiological activity between 120 and 170 ms (a). This is revealed by the strong correlations within this time interval. The histogram shows that the majority of the trials is significantly correlated (b).

disturbing factors, the signal to noise ratio in the EEG measurements is very low. Therefore, the trials are repeated many times and only the average of these trials is then used to identify the ERPs [121]. However, significant information about ERPs is lost, when looking only at trial averages.

In the following, we demonstrate how our visual analytics approach was used to compare entire ensembles of EEG measurements to test the assumption that the same facial stimulus causes the same electrophysiological response in different humans. The analysis considered EEG data of an experiment in which subjects were presented with different variations of a face [121]. Two subjects were randomly selected from this experiment and the measurements from a single electrode were used to compare the interpersonal electrophysiological activity. For the first subject 66 trials and for the second subject 51 trials were available (Figure V.7). The presented stimulus usually evokes a negative potential change at 170 ms, a so called N170. Analyzing the N170 provides important information about a person's sensitivity to faces. Note that while in panel (b) of Figure V.7 a prototypical N170 response can be observed, this is not obvious in panel (a).

To analyze whether both subjects show similar sensitivity to faces, the pairwise WCC between their respective EEG ensembles was computed. Based on their expert knowledge, scientists considered the following parameters to be appropriate: window size of 55 ms, window overlap of 45 ms, and lag range of -25 to 25 ms. The correlations view revealed strong correlations between the ensembles in the time interval between 120 and 170 ms (Figure V.8a). Inspecting individual window-lag combinations with the on-demand histogram showed that within the interval between 120 and 170 ms, the majority of the trials in the ensembles were positively correlated (Figure V.8b).

Our approach clearly reveals that the subjects' responses to faces are similar. Moreover, the high correlation values between 120 and 170 ms show a tendency to be lagged by 10 ms. This suggests that subject one has a slightly slower reaction time than subject two. These insights cannot be readily inferred from the plots shown in Figure V.7 or by only comparing individual trials. From these results it can be concluded that the same facial stimulus causes very similar electrophysiological activity in the two subjects.

## 5.2 Replication of paleoclimate variation derived from stalagmites

In a second example we focus on an important problem in paleoclimatology, where proxy records (such as time series derived from ice cores or stalagmites) from almost the same location would be expected to represent a similar behavior. This is called replication of proxy records and is often not the case, because either the proxies are not reflecting the paleoclimate variation or external factors dominate the climate signal in the proxy record [62, 84].

The comparison involves two proxy records derived from stalagmites collected in Heshang cave [62] and Sanbao cave [147], both located in China. Both records cover the period between 9000 years before present (BP) and 500 years BP. The caves are quite close (approximately 150 km) and the two proxy records should reproduce the same climate signal. The dating of the proxy records involves a certain amount of uncertainty. Therefore, a Monte Carlo approach was used to create ensembles of possible realizations of time series [21].

To analyze the replication, the WCC between both proxy record ensembles was envestigated with our visual analytics approach. A typical time scale of interest to paleoclimatologists is 500 years, with an overlap of 430 years. Since the dating uncertainty in these particular ensembles is up to 200 years, a lag range of ±200 years was chosen. Our tool reveals correlations between both ensembles for the entire time period (Figure V.9). The varying magnitude and uncertainty of correlations uncovers significant variation in the ensembles. This variation, which is also visible in the time series


**Figure V.9:** The windowed cross-correlation between two ensembles of paleoclimate time series. In accordance with domain-specific conventions, the most recent observations are plotted on the left and the oldest on the right.

view (Figure V.9 top), is caused by dating uncertainties. Studying the fractions of significant positive correlations over all combinations of time window and lag (Figure V.10) reveals the generally strong correlation between both ensembles. This particular visualization provides valuable information that allows for reducing the dating uncertainties by extracting a correction function that aligns both proxy records to the same chronology. Note that for some epochs, e.g., around 7700 years BP, the fraction of significant correlations is high over a range of lags (dashed rectangle in Figure V.10). To identify a suitable lag for a correction function, the distributions at the corresponding window-lag combinations were compared. The most suitable lag is the one that has the largest number of high positive correlation values. For example, for the time period at 7700 years BP, the highest correlation values can be found at lag −80 (Figure V.10(b)), whereas at other lags the number of high, though still significant, correlation values is reduced (Figure V.10(c, d)).

From these findings it can be concluded that both proxy records replicate well, although not

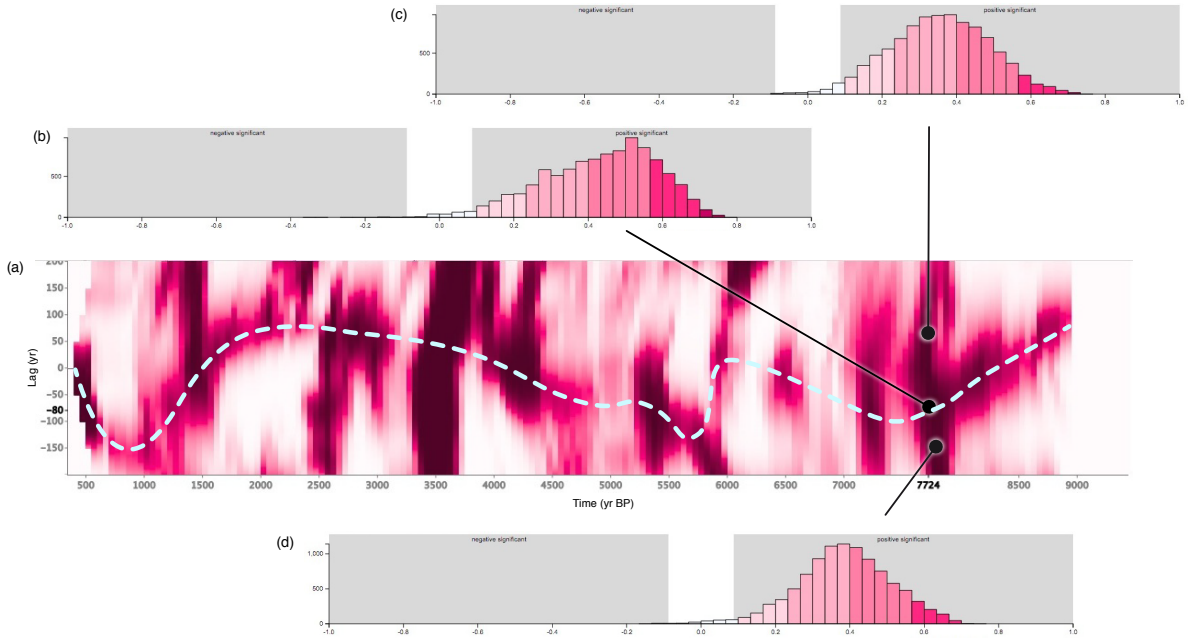



**Figure V.10:** (a) Fractions of significant positive correlations between two ensembles of paleoclimate time series. The dashed line indicates the derived correction function for reducing the dating uncertainties. (b–d) Looking at the distributions of correlation values for different lags at the epoch around 7700 years BP (dashed rectangle) allows for identifying the most suitable lag for the correction function. Since (b) exhibits the largest portion of high correlations, its respective lag was chosen.

perfectly. Main issues are differences between the records caused by unresolved processes influencing the dating procedure. However, the proposed visual analytics approach enabled the extraction of a correction function which reduces the uncertainties in the dating procedure.

## 6 Summary and conclusion

In this paper, we presented a visual analytics approach that extends an established method for correlation-based comparison of two time series – windowed cross-correlation – to support the comparison of entire ensembles of time series. To meet the requirements of time series analysts, we combined semiautomatic statistical analysis with visual exploration in a field-ready visual analytics system. We further build on Hinton diagrams to derive a novel visualization of windowed cross-correlations between ensembles. This matrix-like visualization is visually scalable through semantic zooming and provides an overview of the magnitude and uncertainty of the time-varying correlations between two ensembles. Two use cases demonstrated that our concept allows for gaining valuable insight into

the interrelations between ensembles of time series. Insight which, according to the co-authoring domain experts, could only be obtained with our approach.

Regarding the performance of our system, one has to consider two aspects: the computation of the WCC and the statistical measures, and the interactive visual exploration. The run time of the former highly depends on the application scenario, i.e., the number of time series, the number of observations per time series, and the parameters of the WCC. In the two use cases (Section 5), the WCC calculation took 3 and 37 seconds, respectively (2.6 GHz Intel Core i5 laptop). Since our tool is for domain experts who typically know the appropriate parameters, the computational cost of the WCC calculation was not a concern in our scenarios. The subsequent visual analysis is highly interactive, allowing for real-time exploration of the computed correlations.

To address an even wider range of analysis tasks, we plan two extensions of our approach. (1) Since our concept is also applicable to other methods for studying interrelations between time series, e.g., cross and joint recurrence [49, 99], we intend to incorporate these methods. (2) We would also like to support the comparison of more than two ensembles, adapting our concept to the visual analysis challenges posed by a multi-way comparison.

## Acknowledgements

This research was in part funded by the German Federal Ministry of Education and Research (BMBF project PROGRESS, 03IS2191B).

# Chapter VI

# Synthesis

# 1 Summary

The overall aim of this thesis was to study how visual analytics can facilitate the analysis of processes in geoscientific spatiotemporal data. A thorough analysis of the various processes within the Earth system leads to a better understanding of the complex mechanisms of our planet, e.g., the influence of interactions between the atmosphere and the ocean on global warming [80].

In order to study processes in simulated and observed spatiotemporal data, the geospatial and temporal variability in the data must be analyzed. This task poses several challenges: large spatiotemporal data spaces have to be taken into account, only little aggregation and dimensionality reduction techniques should be applied to reduce loss of information, the most prominent spatiotemporal patterns must be detected, and the detected patterns must be interpreted by domain experts to identify the ones related to processes within the Earth system.

This thesis investigated three research questions, each in a separate chapter, to explore how visual analytics can help address these challenges and advance the analysis of processes in spatiotemporal data. Each research question was studied in an exemplary use case for which a visual analytics solution was developed. The answers to the research questions are as follows:

**Research question 1: How can visual analytics support the detection and assessment of prominent types of spatial situations?** In Chapter III, interactive visual summaries were introduced as a means to facilitate exploration of gridded spatiotemporal data. A visual summary is the depiction of a set of prominent spatial situations in the data and their associated time steps or intervals. It captures the spatial and temporal variability in the data in a compact visual representation. A visual summary also includes interactive means to allow users to assess how well the depicted patterns characterize the original data and to refine the summary where necessary.

In order to enable users to create a visual summary, a visual analytics tool was developed. It uses hierarchical clustering to aggregate all spatial situations in the data into a hierarchy of clusters; a cluster being a set of similar spatial situations. For each cluster, a representative spatial situation is computed. A visual interface enables users to interactively explore and alter this hierarchy, to extract different sets of representative spatial situations, and to assess the corresponding spatiotemporal patterns in a visual summary.

In a use case in ocean modeling and Earth system modeling, interactive visual summaries successfully facilitated the identification and detailed differentiation of El Niño events in satellite obser-

vations of sea surface temperatures.

The results show that clustering is a powerful approach to the detection of prominent types of spatial situations in spatiotemporal data. It enables scientists to take the entire data space into account and does not require them to restrict the analysis to particular geospatial or temporal subsets. Chapter III also demonstrated, that the combination with interactive visual exploration is essential to detect those patterns that are most relevant to the analysis task, to interpret and assess them, and to refine the analytical result where necessary.

Furthermore, the use case illustrated that a visual analytics tool, when specifically tailored to the users' requirements, can complement and sometimes replace established but time-consuming analyses. For example, geoscientists normally combine a variety of methods to detect the mentioned El Niño events, such as regression analysis, empirical orthogonal functions, and wavelet analysis [69]. They also often focus their analysis on various indices that describe particular geographic regions with respect to a specific environmental process [154]. In contrast, the interactive visual summaries approach makes very few assumptions about the data and allows scientists to analyze a variety of patterns and processes for large geographic regions without having to refer to assumption-laden indices.

**Research question 2: How can visual analytics support the detection and assessment of prominent types of temporal behavior?** Chapter IV presents a visual analytics solution that enables ocean modelers to detect geographic regions which exhibit similar temporal behavior, and to use this information in the comparison of ocean model output with reference data.

Previously, ocean modelers relied heavily on strong aggregation of the time dimension, and performed a detailed comparison of the two data sets only for a limited number of hand-picked geographic locations. This strategy lead to significant loss of information and did not provide ocean modelers with a comprehensive overview of differences and similarities between model output and reference data.

The developed visual analytics approach significantly broadens the scope of the analysis by combining an ensemble learning technique – cluster ensembles [131] – with interactive visual exploration. One component enables modelers to perform multiple clusterings of the temporal profiles in model data and reference data. For the clusterings, scientists can use a broad range of parameterizations and features (e.g., descriptive statistics) to take different characteristics of temporal behavior into account and, hence, systematically detect various types of temporal behavior. The resulting two

sets of clusterings represent various perspectives on the temporal behavior in the data. To enable modelers to compare model data and reference data based on these two sets of clusterings, each set is combined into a separate consolidated clustering using cluster ensembles [131]. From a modeler's perspective, the consolidation of clusterings that are based on different features presents a more comprehensive, more robust view on the temporal behavior in a data set. An interactive visual interface allows modelers to subsequently explore, interpret, and compare the two consolidated clusterings of model data and reference data. Since clustering considers the entire geographic domain and systematically detects groups of similar temporal profiles, modelers can now compare geographic regions (the clusters) instead of a few geographic coordinates. This provides them with more options for the detection of processes.

In an ocean modeling use case, the proposed visual analytics approach was successfully applied to the assessment of the Ocean Model for Circulation and Tides (OMCT). The results demonstrate that the integration of user-specified clusterings, consolidation via cluster ensembles, and interactive visual exploration in a visual analytics system is a promising approach to the analysis of prominent types of temporal behavior, and the comparison of two data sets in particular. It also became clear that visual analytics is a valuable complement to modelers' existing tools and routines, with great potential to speed up the model development process.

**Research question 3: How can visual analytics improve the analysis of interrelations of temporal behavior?** Chapter V introduces a visual analytics solution for the detection of interrelations of temporal behavior between sets of time series, in particular ensembles. The approach extends an established technique for the comparison of two individual time series – windowed cross-correlation (WCC) [4, 20] – to the comparison of entire ensembles of time series. To this end, the developed visual analytics system enables geoscientists to compute the cross-correlation between all pairwise combinations of time series from the two ensembles. Exploration and assessment of the complex result regarding interrelations between the two ensembles is facilitated by a visual interface comprising multiple linked views. The core of the visual interface is a novel, glyph-based, matrix-like overview visualization of the WCC.

The proposed approach was employed in two use case. The aim of the first was to study interrelations of temporal behavior between paleoclimatological processes observed in two separate locations in China. In the second use case, electroencephalogram (EEG) scans from two different subjects were compared to find interrelations regarding their response to a particular visual stimulus. In both

cases, the developed visual analytics system enabled scientists to detect and assess interrelations of temporal behavior and to derive valuable scientific insight. According to the collaborating domain experts, this insight could only be obtained with visual analytics because it presents them with new perspectives and new opportunities regarding the comparison of time series ensembles. By extending the established WCC method to the comparison of entire ensembles of time series, visual analytics enabled scientists not only to study interrelations, but also to assess how much these interrelations vary between two ensembles. The latter proved to be especially important to researchers. On the one hand, it prevented them from drawing overconfident conclusions; something that easily happens when applying the WCC to two individual time series only. On the other hand, it allowed scientists to identify significant interrelations between two ensembles even if there was strong variation in the ensembles.

## 2 Main conclusions

This thesis demonstrates that visual analytics is a valuable approach to the analysis of processes in geoscientific spatiotemporal data. Three visual analytics solutions – one for each of the three analysis perspectives outlined in Chapter I, Section 1 – were introduced in Chapters III through V and successfully employed in geoscientific use cases.

Each solution integrates clustering techniques and/or statistical analyses with interactive visual exploration to address the four challenges that were identified as being of particular importance in the analysis of processes in geoscientific spatiotemporal data (see challenges A through D in Chapter I, Section 1). Instead of having to focus only on particular subsets of the data space, scientists are now able to consider the entire data space in the analysis (challenge A), be it through clustering or pairwise computation of windowed cross-correlations between data sets. The analysis is not limited to specific geographic regions, time periods, or time series anymore. Furthermore, much less prior aggregation is now required to detect patterns in spatiotemporal data (challenge B). For example, geoscientists are now able to consider a multitude of features in the detection of prominent types of temporal behavior. Regarding the analysis of interrelations of temporal behavior, no prior aggregation is required at all. As a result of successfully addressing challenges A and B, much less assumptions about the data, and therefore less a priori knowledge, are required at the beginning of the analysis. In addition, researchers are now also better able to detect the most prominent patterns (challenge C), either algorithmically with cluster ensembles or windowed cross-correlation, or via

a combination of hierarchical clustering and interactive visual exploration. Finally, all developed visual analytics solutions have shown that the tight integration of automated analysis with interactive visual exploration facilitates the interpretation of detected patterns (challenge D). Multiple linked views enable geoscientists to harness their expert knowledge to identify the patterns that are most relevant for a given analysis task, and to interpret and assess these patterns for a better understanding of Earth system processes.

This thesis clearly shows that the value of visual analytics for the geosciences is at least threefold. First, visual analytics introduces geoscientists to new, insightful perspectives on the data and the processes they describe. For example, interactive visual summaries (Chapter III) present a novel, concise, and intuitive overview of prominent spatial situations and their occurrence over time in spatiotemporal data. Likewise, the proposed approach for detection of interrelations between ensembles of time series (Chapter V) provides a novel visualization of correlation patterns. This visualization additionally enables assessment of the uncertainty of interrelations between ensembles. The latter example also highlights the second important conclusion that can be drawn from the presented results: Visual analytics typically extends the boundaries of existing analysis methods. Previously, uncertainties regarding the interrelations of temporal behavior between ensembles could only be considered implicitly by comparing the median time series. The proposed visual analytics approach lifts this limitation. It supports an explicit and comprehensive analysis of uncertainties in the interrelations between two ensembles. Similarly, the solution presented in Chapter IV extends existing analysis approaches to the detection of prominent types of temporal behavior by allowing geoscientists to consider various characteristics of temporal behavior. It also constitutes a novel approach to the assessment of simulation model output that is less reliant on a priori knowledge. Finally, this thesis shows that geoscientific applications greatly benefit from the interactivity of visual analytics tools. Researchers are able to interpret and assess the data and the patterns detected via automated analysis in their geographic and temporal context, to conduct visual queries, and to filter as well as to refine the analytical results. Such a free-flowing analytical discourse with the data allows geoscientists to operate at the speed of thought [34] and to receive immediate answers to questions that occur during interactive exploration. This encourages them to follow up on spontaneous ideas or to make educated guesses in the analysis process. Visual analytics, thus, not only facilitates assessment of hypotheses but also generation of new hypotheses about processes in geoscientific spatiotemporal data.

Note that in order to create the visual analytics tools presented in this thesis, a user- and task-

based methodology was imperative (see Chapter II, Section 2). A thorough understanding had to be gained regarding the peculiarities of complex, and often interrelated, real-world phenomena as well as the geoscientific data used to examine them. Furthermore, it was particularly difficult to identify the involved analysis tasks and associated challenges since geoscientists use a significant amount of tacit knowledge and intuition in their analysis. The applied methodology [37] helped to make the tacit knowledge explicit and to obtain the required information. By focusing on the users' needs, this thesis was not only able to propose effective visual analytics solutions, but also to provide the visual analytics community with valuable insight into how geoscientists approach the analysis of spatiotemporal data.

The three presented visual analytics solutions were developed in individual collaborations to facilitate specific use cases. They have yet to prove their value in other contexts. However, since the underlying concepts are generic and have a rather wide scope, it can be assumed that similar results will be achieved in other scenarios. Two of the three approaches support detection and assessment of patterns in gridded spatiotemporal data. Besides oceanography, disciplines such as meteorology or climatology also analyze large amounts of gridded spatiotemporal data and would equally benefit from tools that allow for obtaining an overview of prominent spatiotemporal patterns. Furthermore, many disciplines develop and assess simulation models to analyze the Earth system and its processes. The approach introduced in Chapter IV can be applied to these other modeling contexts since it is not specifically bound to ocean modeling. Note that both approaches were designed for data that are somewhat autocorrelated in time or geographic space. Without autocorrelation, the visualizations of the clustering results would appear noisy and, therefore, be difficult to interpret. In contrast, the visual analytics solution presented in Chapter V does not require autocorrelated data. In fact, it is neither limited to geoscientific data nor ensembles, and theoretically can be used to analyze interrelations between any two sets of time series. Adapting and further developing the proposed solutions for other scenarios and geoscientific applications is an interesting opportunity for future research.

Each visual analytics solution developed in this thesis supports a particular analysis perspective. Since multiple analysis perspectives may be relevant in the same application, the individual solutions can also be used in combination or as modules within a general concept. To pick up on the El Niño example from Chapter I, researchers may first distinguish different states of El Niño in geographic space [136] by detecting the corresponding types of temporal behavior in the data. Next, they can analyze geographic regions associated with particular El Niño states to understand how they

evolve over time. Finally, scientists can compare sets of temporal profiles from different El Niño regions with regard to interrelations of temporal behavior between the different phases. Looking at the data from these three important perspectives enables scientists to gain a more complete picture of spatiotemporal patterns in the data, and, eventually, a more comprehensive understanding of processes. The integration of the three approaches into a general system that blends seamlessly into geoscientists' existing analysis workflow and tool chains is another opportunity for future research.

The requirements and tasks elicited in each collaboration, the identified main analysis perspectives, and the respective visual analytics solutions are a first step towards a general visual analytics solution for analysis of processes in geoscientific data. In order to have many geoscientific disciplines benefit from such a general approach, it would have to support a wide range of analysis questions and perspectives, tasks, and data types. To approach this ambitious endeavor, a general task model for analysis of processes in geoscientific data should be constructed in future research. This requires identification of additional analysis perspectives that are relevant to geoscientists, as well as elicitation of the related tasks and requirements. These findings can then be integrated with the tasks and requirements identified in this thesis. Such a task model would not only be the starting point for the development of a general system, it would also provide guidance for individual visual analytics solutions in a geoscientific context.

In the particular use cases presented in this thesis, computational complexity and visual scalability were not a major issue. However, both points should be considered in future work. Especially, since the amount of data available to geoscientists is growing each year. To keep up with this trend, data models and algorithms that handle increasingly large and heterogeneous geoscientific spatiotemporal data will have to be developed. Extremely large amounts of spatiotemporal data will also pose significant challenges for interactive visual analysis. One promising approach in this regard is *progressive visual analytics* [130], which enables users to explore partial results and interact with the automated analysis to prioritize subspaces that exhibit promising patterns. It is a future research question how this concept can be adopted in geoscientific scenarios. In addition to studying novel ways of interacting with the automated analysis part, scalable visualization techniques will also have to be developed in future research. These techniques should preserve the all-important spatiotemporal context without placing a heavy perceptual burden on users during visual analysis of large spatiotemporal data.

# References


[1] J. Ahrens, K. Heitmann, M. Petersen, J. Woodring, S. Williams, P. Fasel, C. Ahrens, C.-H. Hsu, and B. Geveci. Verifying scientific simulations via comparative and quantitative visualization. *IEEE Computer Graphics and Applications*, 30(6):16–28, 2010.

[2] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of time-oriented data*. Springer, 2011.

[3] R. Allan, J. Lindesay, and D. Parker. El Niño Southern Oscillation and climatic variability. *Oceanographic Literature Review*, 44(6):555, 1997.

[4] U. Altmann. Investigation of movement synchrony using windowed cross-lagged regression. In A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt, editors, *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, volume 6800 of *LNCS*, pages 335–345. Springer, Berlin Heidelberg, 2011.

[5] R. Amar and J. Stasko. A knowledge task-based framework for design and evaluation of information visualizations. In *Proceedings IEEE Symposium on Information Visualization 2004*, pages 143–149, Washington, DC, 2004. IEEE.

[6] G. Andrienko, N. Andrienko, P. Bak, S. Bremm, D. Keim, T. von Landesberger, C. Pölitz, and T. Schreck. A framework for using self-organising maps to analyse spatio-temporal patterns, exemplified by analysis of mobile phone usage. *Journal of Location Based Services*, 4(3–4):200–221, 2010.

[7] G. Andrienko, N. Andrienko, S. Bremm, T. Schreck, T. Von Landesberger, P. Bak, and D. Keim. Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. *Computer Graphics Forum*, 29(3):913–922, 2010.

[8] G. Andrienko, N. Andrienko, J. Dykes, D. Mountain, P. Noy, M. Gahegan, J. C. Roberts, P. Rodgers, and M. Theus. Creating instruments for ideation: Software approaches to geovisualization. In J. Dykes, A. M. MacEachren, and M.-J. Kraak, editors, *Exploring Geovisualization*, chapter 5, pages 103–125. Elsevier, 2005.

[9] N. Andrienko and G. Andrienko. *Exploratory analysis of spatial and temporal data*. Springer, Berlin, Heidelberg, 2006.

[10] N. Andrienko and G. Andrienko. A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery*, 27(1):55–83, 2013.

[11] N. Andrienko, G. Andrienko, and P. Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, 14(6):503–541, 2003.

[12] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007.

[13] E. Başar, C. Başar-Eroglu, R. Parnefjord, E. Rahn, and M. Schürmann. Evoked potentials: Ensembles of brain induced rhythmicities in the alpha, theta and gamma ranges. In E. Başar and T. H. Bullock, editors, *Induced Rhythms in the Brain*, Brain Dynamics, pages 155–181. Birkhäuser, Boston, 1992.

[14] C. Bennett, P. Gacs, M. Li, P. Vitanyi, and W. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.

[15] P. Berkhin. A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, and M. Teboulle, editors, *Grouping Multidimensional Data*, pages 25–71. Springer, Berlin, Heidelberg, 2006.

[16] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *AAAI-94 Workshop on Knowledge Discovery in Databases*, volume 10, pages 359–370. Seattle, WA, 1994.

[17] L. Berry and T. Munzner. Binx: Dynamic exploration of time series datasets across aggregation levels. In *IEEE Symposium on Information Visualization, 2004.* IEEE, Oct 2004.

[18] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps.* University of Wisconsin Press, 1983.

[19] C. Blok. Monitoring change: characteristics of dynamic geo-spatial phenomena for visual exploration. In C. Freksa, W. Brauer, C. Habel, and K. F. Wender, editors, *Spatial Cognition II*, volume 1849 of *Lecture Notes in Computer Science*, pages 16–30. Springer, 2000.

[20] S. M. Boker, J. L. Rotondo, M. Xu, and K. King. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychol. Methods*, 7(3):338–355, 2002.

[21] S. F. M. Breitenbach, K. Rehfeld, B. Goswami, J. U. L. Baldini, H. E. Ridley, D. Kennett, K. Prufer, V. V. Aquino, Y. Asmerom, V. J. Polyak, H. Cheng, J. Kurths, and N. Marwan. Constructing Proxy-Record Age models (COPRA). *Clim. Past*, 8:1765–1779, 2012.

[22] F. J. Bremner, S. J. Gotts, and D. L. Denham. Hinton diagrams: Viewing connection strengths in neural networks. *Behav. Res. Meth. Ins. C.*, 26(2):215–218, 1994.

[23] S. Bruckner and T. Möller. Result-Driven Exploration of Simulation Parameter Spaces for Visual Effects Design. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1468–1476, 2010.

[24] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think.* Morgan Kaufmann, 1999.

[25] S. M. Casner. A task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics*, 10(2):111–151, 1991.

[26] A. Cedilnik and P. Rheingans. Procedural annotation of uncertain information. In T. Ertl, B. Hamann, and A. Varshney, editors, *Visualization 2000: October 8 - 13, 2000, Salt Lake City, UT, USA*, pages 77–83. IEEE, Piscataway, NJ, USA, 2000.

[27] C. Chen. *Information visualization: Beyond the horizon.* Springer Science & Business Media, 2006.

[28] J. Chen, A. M. MacEachren, and D. J. Peuquet. Constructing Overview + Detail Dendrogram-Matrix Views. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):889–896, 2009.

[29] M. Cover Thomas and A. Thomas Joy. *Elements of information theory*. Wiley, New York, 1991.

[30] B. Craft and P. Cairns. Beyond guidelines: what can we learn from the visual information seeking mantra? In *Proceedings 9th International Conference on Information Visualisation, 2005*, pages 110–118, July 2005.

[31] R. Dachselt, M. Frisch, and M. Weiland. Facetzoom: A continuous multi-scale widget for navigating hierarchical metadata. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, CHI '08, pages 1353–1356, New York, NY, USA, 2008. ACM.

[32] L. De Grandis. *Theory and use of color*. Abrams, 1986.

[33] U. Demšar and K. Virrantaus. Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science*, 24(10):1527–1542, 2010.

[34] J. E. Devaney, S. Satterfield, J. Hagedorn, J. Kelso, A. Peskin, W. George, T. Griffin, H. Hung, and R. Kriz. Science at the speed of thought. In Y. Cai, editor, *Ambient Intelligence for Scientific Discovery*, volume 3345 of *Lecture Notes in Computer Science*, pages 1–24. Springer, Berlin Heidelberg, 2005.

[35] Diansheng Guo, Jin Chen, A. M. MacEachren, and Ke Liao. A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1461–1474, 2006.

[36] M. Dodge, M. McDerby, and M. Turner, editors. *Geographic visualization*. John Wiley & Sons, Chichester, West Sussex, England, 2008.

[37] D. Dransch, P. Köthur, S. Schulte, V. Klemann, and H. Dobslaw. Assessing the quality of geoscientific simulation models with visual analytics methods – a design study. *International Journal of Geographical Information Science*, 24(10):1459–1479, 2010.

[38] Y. Drocourt, R. Borgo, K. Scharrer, T. Murray, S. Bevan, and M. Chen. Temporal visualization of boundary-based geo-information using radial projection. *Computer Graphics Forum*, 30(3):981–990, 2011.

[39] J. Dykes, A. M. MacEachren, and M.-J. Kraak, editors. *Exploring geovisualization*. Elsevier, Oxford, 2005.

[40] N. Elmqvist, T.-N. Do, H. Goodell, N. Henry, and J.-D. Fekete. Zame: Interactive large-scale graph visualization. In *Proc. PacificVIS '08*, pages 215–222. IEEE, March 2008.

[41] N. Elmqvist and J.-D. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE T. Vis. Comput. Gr.*, 16(3):439–454, May 2010.

[42] D. B. Enfield and D. A. Mayer. Tropical atlantic sea surface temperature variability and its relation to El Niño-Southern Oscillation. *Journal of Geophysical Research: Oceans*, 102(C1):929–945, 1997.

[43] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231, 1996.

[44] C. Fish, K. P. Goldsberry, and S. Battersby. Change Blindness in Animated Choropleth Maps: An Empirical Study. *Cartography and Geographic Information Science*, 38(4):350–362, 2011.

[45] P. Fox and J. Hendler. Changing the equation on scientific data visualization. *Science*, 311:705–708, 11 Feb 2011.

[46] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[47] A. L. N. Fred and A. Jain. Data clustering using evidence accumulation. In *Proceedings of the 16th International Conference on Pattern Recognition*, volume 4, pages 276–280, 2002.

[48] S. Frey, F. Sadlo, and T. Ertl. Visualization of temporal similarity in field data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2023–2032, 2012.

[49] B. Goswami, N. Marwan, G. Feulner, and J. Kurths. How do global temperature drivers influence each other? – A network perspective using recurrences. *Eur. Phys. J. – Special Topics*, 222:861–873, 2013.

[50] J. Han and M. Kamber. *Data mining: Concepts and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2006.

[51] M. C. Hao, M. Marwah, H. Janetzko, U. Dayal, D. A. Keim, D. Patnaik, N. Ramakrishnan, and R. K. Sharma. Visual exploration of frequent patterns in multivariate time series. *Information Visualization*, 11(1):71–83, 2012.

[52] M. Harrower and C. A. Brewer. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[53] M. Hashizume, T. Terao, and N. Minakawa. The Indian Ocean Dipole and malaria risk in the highlands of western Kenya. *Proc. NAS USA*, 106(6):1857–62, 2009.

[54] I. Herman, G. Melancon, and M. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.

[55] B. C. Hewitson. Climate Analysis, Modelling, and Regional Downscaling Using Self-Organizing Maps. In P. Agarwal and A. Skupin, editors, *Self-Organizing Maps: Applications in Geographic Information Science*, pages 137–153. John Wiley & Sons, Ltd and Wiley, Chichester and UK, 2008.

[56] T. Hey, S. Tansley, and K. Tolle, editors. *The fourth paradigm: Data-intensive scientific discovery*, Redmond, WA, 2009. Microsoft Research.

[57] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. Distributed representations. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 77–109. MIT Press, Cambrige, MA, 1986.

[58] G. E. Hinton and T. J. Sejnowski. Learning and relearning in boltzmann machines. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 282–317. MIT Press, Cambrige, MA, 1986.

[59] T. Höllt, A. Magdy, P. Zhan, G. Chen, G. Gopalakrishnan, I. Hoteit, C. D. Hansen, and M. Hadwiger. Ovis: A framework for visual analysisof ocean forecast ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 20(8):1114–1126, 2014.

[60] I. Horenko. On clustering of non-stationary meteorological time series. *Dynamics of Atmospheres and Oceans*, 49(2-3):164–187, 2010.

[61] K.-C. Hsu and S.-T. Li. Clustering spatial–temporal precipitation data using wavelet transform and self-organizing map neural network. *Advances in Water Resources*, 33(2):190–200, 2010.

[62] C. Hu, G. M. Henderson, J. Huang, S. Xie, Y. Sun, and K. R. Johnson. Quantification of Holocene Asian monsoon rainfall from spatially separated cave records. *Earth Planet. Sc. Lett.*, 266(3-4):221–232, 2008.

[63] R. Huth. An intercomparison of computer-assisted circulation classification methods. *International Journal of Climatology*, 16(8):893–922, 1996.

[64] R. Huth. A circulation classification scheme applicable in GCM studies. *Theoretical and Applied Climatology*, 67(1-2):1–18, 2000.

[65] R. F. Hüttl, editor. *Ein Planet voller Überraschungen/Our Surprising Planet: Neue Einblicke in das System Erde/New Insights into System Earth*. Spektrum Akademischer Verlag, Heidelberg, 2011.

[66] N. Iam-On, S. Garrett, C. Price, and T. Boongoen. Link-based cluster ensembles for heterogeneous biological data analysis. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2010*, pages 573–578, Dec 2010.

[67] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[68] D. Kao, J. L. Dungan, and A. T. Pang. Visualizing 2D Probability Distributions from EOS Satellite Image-Derived Data Sets: A Case Study. In *Visualization 2001: 21-26 October, San Diego, CA, USA*, pages 457–460. IEEE, 2001.

[69] H.-Y. Kao and J.-Y. Yu. Contrasting eastern-pacific and central-pacific types of enso. *Journal of Climate*, 22(3):615–632, 2014/08/17 2009.

[70] G. Karypis, E.-H. Han, and V. Kumar. Chameleon: hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, Aug 1999.

[71] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[72] J. Kehrer. *Interactive Visual Analysis of Multi-faceted Scientific Data*. PhD thesis, University of Bergen, 2011.

[73] J. Kehrer and H. Hauser. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):495–513, March 2013.

[74] J. Kehrer, F. Ladstädter, P. Muigg, H. Doleisch, A. Steiner, and H. Hauser. Hypothesis Generation in Climate Research with Interactive Visual Data Exploration. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1579–1586, Nov 2008.

[75] J. Kehrer, P. Muigg, H. Doleisch, and H. Hauser. Interactive visual analysis of heterogeneous scientific data across an interface. *IEEE Transactions on Visualization and Computer Graphics*, 17(7):934–946, July 2011.

[76] D. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, Jan 2002.

[77] D. Keim, J. Kohlhammer, G. Ellis, and F. Mannsmann, editors. *Mastering the information age: Solving problems with visual analytics,*. Eurographics Association, Goslar, 2010.

[78] D. A. Keim, F. Mansmann, and J. Thomas. Visual analytics: How much visualization and how much analytics? *SIGKDD Explor. Newsl.*, 11(2):5–8, May 2009.

[79] L. Knapp. A task analysis approach to the visualization of geographic data. In T. L. Nyerges, D. M. Mark, R. Laurini, and M. J. Egenhofer, editors, *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems*, pages 355–371. Kluwer Academic Publishers, 1995.

[80] Y. Kosaka and S.-P. Xie. Recent global-warming hiatus tied to equatorial pacific surface cooling. *Nature*, 501(7467):403–407, 09 2013.

[81] P. Köthur, M. Sips, J. Kuhlmann, and D. Dransch. Visualization of geospatial time series from environmental modeling output. In M. Meyer and T. Weinkauf, editors, *Proceedings of the Eurographics Conference on Visualization (EuroVis) 2012 Short Papers*, pages 115–119, Goslar, 2012. Eurographics Association.

[82] P. Köthur, M. Sips, A. Unger, J. Kuhlmann, and D. Dransch. Interactive visual summaries for detection and assessment of spatiotemporal patterns in geospatial time series. *Information Visualization*, 13(3):283–298, 2014.

[83] M. Kreuseler and H. Schumann. A flexible approach for visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):39–51, 2002.

[84] M. S. Lachniet. Climatic and environmental controls on speleothem oxygen-isotope values. *Quaternary Sci. Rev.*, 28:412–432, 2009.

[85] F. Ladstädter, A. K. Steiner, B. C. Lackner, B. Pirscher, G. Kirchengast, J. Kehrer, H. Hauser, P. Muigg, and H. Doleisch. Exploration of Climate Data Using Interactive Visualization. *Journal of Atmospheric and Oceanic Technology*, 27(4):667–679, 2010.

[86] G. N. Lance and W. T. Williams. A Generalized Sorting Strategy for Computer Classifications. *Nature*, 212(5058):218, 1966.

[87] C. Lange-Küttner. Ebbinghaus simulated: Just do it 200 times. In *IEEE Int. Conf. on Development and Learning (ICDL) 2011*, volume 2, pages 1–6. IEEE, Aug 2011.

[88] T. W. Liao. Clustering of time series data – a survey. *Pattern Recognition*, 38(11):1857 – 1874, 2005.

[89] J. Lin, E. Keogh, and S. Lonardi. Visualizing and discovering non-trivial patterns in large time series databases. *Information Visualization*, 4(2):61–82, 2005.

[90] D. R. Lipşa, R. S. Laramee, S. J. Cox, J. C. Roberts, R. Walker, M. A. Borkin, and H. Pfister. Visualization for the physical sciences. *Computer Graphics Forum*, 31(8):2317–2347, 2012.

[91] Liwei Wang, Yan Zhang, and Jufu Feng. On the Euclidean distance of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1334–1339, 2005.

[92] S. J. Luck. *An introduction to the event-related potential technique*. MIT Press, Cambridge, 2005.

[93] A. M. MacEachren. Visualizing uncertain information. *Cartographic Perspectives*, 13:10–19, 1992.

[94] A. M. MacEachren. *How maps work: representation, visualization, and design*. Guilford Press, 2004.

[95] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005.

[96] H. Madsen. *Time Series Analysis*. Chapman & Hall/CRC texts in statistical science. Chapman & Hall/CRC, Boca Raton, 2008.

[97] A. Malik, R. Maciejewski, S. Towers, S. McCullough, and D. Ebert. Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1863–1872, Dec 2014.

[98] N. Marwan. Windowed cross correlation (corrgram). MATLAB Central File Exchange, June 2007.

[99] N. Marwan and J. Kurths. Nonlinear analysis of bivariate data with cross recurrence plots. *Phys. Lett. A*, 302(5–6):299–307, 2002.

[100] K. Matković, D. Gračanin, M. Jelović, A. Ammer, A. Lež, and H. Hauser. Interactive visual analysis of multiple simulation runs using the simulation model view: Understanding and tuning of an electronic unit injector. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1449–1457, Nov 2010.

[101] K. Mehlhorn. *Sortieren und Suchen*, volume 1 of *Datenstrukturen und effiziente Algorithmen*. Teubner, Stuttgart, 2 edition, 1988.

[102] H. J. Miller and J. Han, editors. *Geographic data mining and knowledge discovery*. Taylor & Francis, London and New York, 2001.

[103] P. Muigg, J. Kehrer, S. Oeltze, H. Piringer, H. Doleisch, B. Preim, and H. Hauser. A four-level focus+context approach to interactive visual analysis of temporal features in large scientific data. *Computer Graphics Forum*, 27(3):775–782, May 2008.

[104] D. Müllner. Modern hierarchical, agglomerative clustering algorithms. `http://arxiv.org/abs/1109.2378v1`, 2011. Accessed: 30 August 2012.

[105] F. Murtagh and A. Heck. *Multivariate Data Analysis*. Kluwer, Dordrecht, 1987.

[106] E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk, and D. Imre. ClusterSculptor: A Visual Analytics Tool for High-Dimensional Data. In *Proceedings of the IEEE Symposium on Visual Analytics and Technology 2007*, pages 75–82, 2007.

[107] M. Nisha, S. Mohanavalli, and R. Swathika. Improving the quality of clustering using cluster ensembles. In *Proceedings of the IEEE Conference on Information Communication Technologies (ICT) 2013*, pages 88–92, 2013.

[108] T. Nocke, M. Flechsig, and U. Bohm. Visual exploration and evaluation of climate-related simulation data. In *Proceedings of Winter Simulation Conference 2007*, pages 703–711, 2007.

[109] M. Novotnỳ and H. Hauser. Outlier-preserving focus+context visualization in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):893–900, Sept 2006.

[110] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha. Approaches to uncertainty visualization. *Visual Computer*, 13(8):370–390, 1997.

[111] D. Peuquet. It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3):441–461, 1994.

[112] H. Piringer, S. Pajer, W. Berger, and H. Teichmann. Comparative visual analysis of 2d function ensembles. *Computer Graphics Forum*, 31(3pt3):1195–1204, 2012.

[113] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. Schwalm, R. Cook, E. Bertini, and C. Silva. SimilarityExplorer: A Visual Inter-Comparison Tool for Multifaceted Climate Data. *Computer Graphics Forum*, 33(3):341–350, 2014.

[114] K. Potter, J. Kniss, R. Riesenfeld, and C. R. Johnson. Visualizing Summary Statistics and Uncertainty. In G. Melançon, T. Munzner, and D. Weiskopf, editors, *Eurographics/IEEE-VGTC Symp. Visualization 2010: 9-11 June 2010, Bordeaux, France*, pages 823–832. Blackwell, Oxfor, UK, 2010.

[115] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johnson. Ensemble-Vis: A Framework for the Statistical Visualization of Ensemble Data. In *Proc. ICDMW '09, 6 Dec 2009, Miami, Florida*, pages 233–240. IEEE, Los Alamitos and CA, 2009.

[116] K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths. Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Proc. Geoph.*, 18(3):389–404, 2011.

[117] S. R. Rintoul, C. Hughes, and D. Olbers. The antarctic circumpolar current system. In J. H. Steele, S. A. Thorpe, and K. K. Turekian, editors, *Ocean Currents: A derivative of the encyclopedia of Ocean Sciences*, chapter 4.6, pages 271–302. Academic Press, London, 2nd edition, 2010.

[118] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[119] H. W. Rust, M. Vrac, M. Lengaigne, and B. Sultan. Quantifying Differences in Circulation Patterns Based on Probabilistic Models: IPCC AR4 Multimodel Comparison for the North Atlantic. *Journal of Climate*, 23(24):6573–6589, 2010.

[120] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. J. Moorhead. Noodles: A Tool for Visualization of Numerical Weather Model Ensemble Uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1421–1430, 2010.

[121] S. Schinkel, G. Ivanova, J. Kurths, and W. Sommer. Modulation of the N170 adaptation profile by higher level factors. *Biol. Psychol.*, 97:27–34, 2014.

[122] H. Schumann and W. Müller. *Visualisierung: Grundlagen und allgemeine Methoden*. Springer, Berlin, Heidelberg, New York, 2000.

[123] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[124] J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results. *Computer*, 35(7):80–86, 2002.

[125] J. Seo and B. Shneiderman. Knowledge discovery in high-dimensional data: case studies and a user survey for the rank-by-feature framework. *IEEE Transactions on Visualization and Computer Graphics*, 12(3):311–322, 2006.

[126] E. I. Shakhnovich and A. M. Gutin. Engineering of stable and fast-folding sequences of model proteins. *Proc. NAS*, 90(15):7195–7199, 1993.

[127] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings of IEEE Symposium on Visual Languages 1996*, pages 336–343. IEEE, 1996.

[128] T. A. Slocum, R. B. McMaster, F. C. Kessler, and H. H. Howard. *Thematic Cartography and Geographic Visualization*. Prentice Hall Series in Geographic Information Science. Prentice Hall, 3rd edition, 2008.

[129] R. Spence. *Information Visualization: An Introduction*. Springer, Chambridge, Heidelberg, New York, Dordrecht, London, 3rd edition, 2014.

[130] C. Stolper, A. Perer, and D. Gotz. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1653–1662, Dec 2014.

[131] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 2002(3):583–617, 2002.

[132] B. D. Tapley, S. Bettadpur, J. C. Ries, P. F. Thompson, and M. M. Watkins. Grace measurements of mass variability in the earth system. *Science*, 305(5683):503–505, 2004.

[133] J. J. Thomas and K. A. Cook. *Illuminating the path: The research and development agenda for visual analytics*. IEEE, Los Alamitos, CA, USA, 1st edition, 2005.

[134] M. Thomas, J. Sündermann, and E. Maier-Reimer. Consideration of ocean tides in an ogcm and impacts on subseasonal to decadal polar motion excitation. *Geophysical Research Letters*, 28(12):2457–2460, 2001.

[135] C. Tominski, J. Donges, and T. Nocke. Information visualization in climate research. In *Proceedings of the 15th International Conference on Information Visualization (IV)*, pages 298–305, 2011.

[136] K. E. Trenberth. The definition of El Niño. *Bulletin of the American Meteorological Society*, 78(12):2771–2777, 1997.

[137] K. E. Trenberth, editor. *Climate System Modeling*. Cambridge Universtity Press, New York, 2009.

[138] E. R. Tufte. *Envisioning information*. Graphics Press, Cheshire and CT, 1990.

[139] E. R. Tufte. *The visual display of quantitative information*. Graphics Press, Cheshire and CT, 2 edition, 2001.

[140] J. W. Tukey. *Exploratory data analysis*. Addison-Wesley, Reading, MA, 1977.

[141] B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4):247–262, 2002.

[142] A. Unger, S. Schulte, V. Klemann, and D. Dransch. A visual analysis concept for the validation of geoscientific simulation models. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2216–2225, 2012.

[143] T. C. Urdan. *Statistics in Plain English*. Routledge, New York, London, 3rd edition, 2010.

[144] T. van Long and L. Linsen. MultiClusterTree: Interactive Visual Exploration of Hierarchical Clusters in Multidimensional Multivariate Data. *Computer Graphics Forum*, 28(3):823–830, 2009.

[145] R. F. P. van Pelt, S. S. A. M. Jacobs, B. M. ter Haar Romeny, and A. Vilanova. Visualization of 4D Blood-Flow Fields by Spatiotemporal Hierarchical Clustering. *Computer Graphics Forum*, 31(3pt2):1065–1074, 2012.

[146] J. J. van Wijk and E. R. van Selow. Cluster and calendar based visualization of time series data. In G. Wills and D. Keim, editors, *Proceedings of the 1999 IEEE Symposium on Information Visualization (InfoVis'99)*, pages 4–9. IEEE, 1999.

[147] Y. Wang, H. Cheng, R. L. Edwards, X. Kong, X. Shao, S. Chen, J. Wu, X. Jiang, X. Wang, and Z. An. Millennial- and orbital-scale changes in the East Asian monsoon over the past 224,000 years. *Nature*, 451(7182):1090–3, 2008.

[148] C. Ware. *Information visualization: perception for design*. Morgan Kaufmann, Waltham, MA, 3rd edition, 2013.

[149] P. J. Webster, G. J. Holland, J. A. Curry, and H.-R. Chang. Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science*, 309(5742):1844–1846, 2005.

[150] S. Wehrend and C. Lewis. A problem-oriented classification of visualization techniques. In *Proceedings of the First IEEE Conference on Visualization: Visualization'90*, pages 139–143. IEEE, 1990.

[151] J. Woodring and H.-W. Shen. Multiscale Time Activity Data Exploration via Temporal Clustering Visualization Spreadsheet. *IEEE Transactions on Visualization and Computer Graphics*, 15(1):123–137, 2009.

[152] H.-M. Wu, S. Tzeng, and C.-h. Chen. Matrix visualization. In C. Chun-houh, W. K. Härdle, and A. Unwin, editors, *Handbook of Data Visualization*, Springer Handbooks Comp. Statistics, pages 681–708. Springer, Berlin, Heidelberg, 2008.

[153] H. Yu, C. Wang, and K.-L. Ma. Parallel hierarchical visualization of large time-varying 3D vector fields. In *Proceedings of the 2007 ACM/IEEE Conference on Supercomputing*, pages 1–12, 2007.

[154] J. Yu and S. Kim. Identifying the types of major El Niño events since 1870. *International Journal of Climatology*, 2012 (online first).

[155] Z. Yu, H.-S. Wongb, J. You, Q. Yang, and H. Liao. Knowledge based cluster ensemble for cancer discovery from biomolecular data. *IEEE Transactions on NanoBioscience*, 10(2):76–85, 2011.

[156] A. Zelenyuk, D. Imre, Y. Cai, K. Mueller, Y. Han, and P. Imrich. SpectraMiner, an interactive data mining and visualization software for single particle mass spectroscopy: A laboratory test case. *International Journal of Mass Spectrometry*, 258(1-3):58–73, 2006.

[157] X. Zhang, L. Jiao, F. Liu, L. Bo, and M. Gong. Spectral clustering ensemble applied to SAR image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 46(7):2126–2136, 2008.

[158] M. X. Zhou and S. K. Feiner. Visual task characterization for automated visual discourse synthesis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 392–399, New York, NY, 1998. ACM / Addison-Wesley.

[159] Z.-H. Zhou. *Ensemble methods: foundations and algorithms*. Machine learning & pattern recognition series. Chapman & Hall/CRC, Boca Raton, FL, 2012.

[160] L. Zubair, G. N. Galappaththy, H. Yang, J. Chandimala, Z. Yahiya, P. Amerasinghe, N. Ward, and S. J. Connor. Epochal changes in the association between malaria epidemics and El Niño in Sri Lanka. *Malaria J.*, 7(1):140, 2008.

[161] T. D. Zuk. *Visualizing Uncertainty*. PhD thesis, University of Calgary, Canada, 2008. AAINR38246.

# Eidesstattliche Erklärung

Hiermit erkläre ich, die vorliegende Dissertation selbstständig und ohne Verwendung unerlaubter Hilfe angefertigt zu haben. Die aus fremden Quellen direkt oder indirekt übernommenen Inhalte sind als solche kenntlich gemacht. Die Dissertation wird erstmalig und nur an der Humboldt Universität zu Berlin eingereicht. Weiterhin erkläre ich, nicht bereits einen Doktortitel im Fach Geographie zu besitzen. Die dem Verfahren zu Grunde liegende Promotionsordnung ist mir bekannt.

Berlin, den 13. Juli 2015 Patrick Köthur